Methods Guide for
Comparative Effectiveness Reviews

Assessing the Applicability of Studies When Comparing Medical Interventions

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

Comparative Effectiveness Reviews are systematic reviews of existing research on the effectiveness, comparative effectiveness, and harms of different health care interventions. They provide syntheses of relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. Strong methodologic approaches to systematic review improve the transparency, consistency, and scientific rigor of these reports. Through a collaborative effort of the Effective Health Care (EHC) Program, the Agency for Healthcare Research and Quality (AHRQ), the EHC Program Scientific Resource Center, and the AHRQ Evidence-based Practice Centers have developed a *Methods Guide for Comparative Effectiveness Reviews*. This Guide presents issues key to the development of Comparative Effectiveness Reviews and describes recommended approaches for addressing difficult, frequently encountered methodological issues.

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

This document was written with support from the Effective Health Care Program at AHRQ.

The views expressed in this paper are those of the authors and do not represent the official policies of the Agency for Healthcare Research and Quality, the Department of Health and Human Services, the Department of Veterans Affairs, the Veterans Health Administration, or the Health Services Research and Development Service.

None of the authors has a financial interest in any of the products discussed in this document.

Suggested citation: Atkins D, Chang S, Gartlehner G, Buckley DI, Whitlock EP, Berliner E, Matchar D. Assessing the Applicability of Studies When Comparing Medical Interventions. Agency for Healthcare Research and Quality; January 2011. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 11-EHC019-EF. Available at http://effectivehealthcare.ahrq.gov/.

Assessing the Applicability of Studies When Comparing Medical Interventions

Authors:

David Atkins, M.D., M.P.H.[1]

Stephanie Chang, M.D., M.P.H.[2]

Gerald Gartlehner, M.D., M.P.H.[3]

David I. Buckley, M.D., M.P.H.[4]

Evelyn P. Whitlock, M.D., M.P.H.[5]

Elise Berliner, Ph.D.[2]

David Matchar, M.D., FACP[6,7]

[1]Office of Research and Development, Department of Veterans Affairs, Washington, DC.
[2]Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD.
[3]Department for Evidence-based Medicine and Clinical Epidemiology, Danube University, Krems, Austria.
[4]Oregon Evidence-based Practice Center, Oregon Health and Science University, Portland, OR.
[5]Center for Health Research, Kaiser Permanente Northwest, Portland, OR.
[6]Duke Center for Clinical Health Policy Research, Durham, NC.
[7]Duke-NUS Medical School, Singapore.

Assessing the Applicability of Studies When Comparing Medical Interventions

Key Points

- The PICOS framework is a useful way of organizing the review and presentation of factors that affect applicability.
- Input from clinical experts and stakeholders can help identify specific study elements that should be routinely abstracted to examine applicability.
- Population-based surveys, pharmacoepidemiologic studies, and large case series or registries of devices or surgical procedures can be used to determine whether the populations, interventions, and comparisons in existing studies are representative of current practice.
- Reviewers should assess whether benefits or harms vary along with differences in patient or intervention characteristics (i.e. effect modification) or with differences in underlying risk.
- Reports should clearly highlight important issues relevant to applicability of individual studies in a "Comments" or "Limitations" section of evidence tables and in text.
- Meta-regression, sub-group analysis and/or separate applicability summary tables may help reviewers and those using the reports see how well the body of evidence applies to the question at hand.
- Judgments about applicability of the evidence should consider the entire body of studies.

Introduction

A defining characteristic of comparative effectiveness research is that it includes "the conduct and synthesis of research comparing the benefits and harms of different interventions… in 'real world' settings" with the purpose of determining "which interventions are most effective for which patients under specific circumstances."[1] A comparative effectiveness review must therefore make judgments about whether the available research evidence reflects "real world" practice and should make clear for which patients and which circumstances the review's conclusions can be used to make clinical or policy decisions. Existing guidance on conducting systematic reviews has focused on the risk of bias in individual studies and judging whether conclusions of the review are internally valid, rather than this equally important aspect of the review process.[2]

A variety of terms have been used to describe this aspect—*applicability, external validity, generalizability, directness,* and *relevance*. Shadish and Cook define *external validity* as "inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments and outcomes."[3] The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group has used the term *directness* to cover applicability as well as other distinct aspects of the relationship between the evidence and making recommendations[4]. We prefer *applicability*, which we define as the extent to which the effects observed in published studies are likely to reflect the expected results when a specific intervention is applied to the population of interest under "real-world" conditions. This better reflects the perspective of reviews conducted by the Agency for Healthcare Research and Quality

(AHRQ) Effective Health Care (EHC) Program and by many other groups (for example, guideline developers) in which systematic review aim to answer specific clinical or policy questions involving particular populations and then must make judgments about whether the available evidence is *applicable* to the questions at hand.

Relatively few clinical trials are designed with applicability in mind and furthermore, clinical studies typically report only a few of the factors needed to fully assess applicability. In contrast to the accumulating body of empiric data on factors affecting the risk of bias, or internal validity, there has been much less empiric data to determine which factors affect applicability. For these reasons, to date there has not been any detailed guidance for assessing applicability of evidence in producing systematic reviews.

This paper outlines specific steps to ensure that systematic reviews describe and characterize the evidence so that users of a review can apply it appropriately in their decisions. The first step, identifying factors that may affect applicability, should be considered at the very earliest stages of a review, when defining key questions and the populations, interventions, comparators, and outcomes of interest. Defining inclusion and exclusion criteria inevitably takes into account factors that may affect the applicability of studies—for example, reviews meant to inform decision-makers in developed countries exclude studies in developing countries because they may not be applicable to the patients and health care settings in Western countries. This paper focuses on subsequent steps in a review to describe a systematic but practical approach for considering applicability in the process of reviewing, reporting, and synthesizing evidence from eligible studies.

To develop this guidance, we searched the literature using the terms *applicability* and *external validity* and reviewed our own experience with working with users of reviews produced by the Evidence-based Practice Center (EPC) program. We extracted specific study characteristics which were proposed as relevant to external validity or applicability in the literature; the paper of Rothwell[5] provided an extensive list to which we added from other literature, prioritized based on the experience of our program, and organized under the PICOS framework (Patient, Intervention, Comparator, Outcome, Setting). We presented draft guidance at in-person meetings of the EPC program and circulated multiple drafts for review by EPC investigators. Parts of an earlier draft were posted for public comment. The final guidance document has incorporated peer and public review comments.

General Guidance

Applicability Should Be Judged Separately for Different Outcomes

The most applicable evidence may differ when considering benefits or harms since these often depend on distinct physiologic processes. For example, evidence of the benefits of aspirin for prevention of cardiovascular events from patients with heart disease cannot be readily applied to healthy populations. However, studies of patients with and without heart disease may be useful for estimating the gastrointestinal risks of aspirin which act through different mechanisms and do not vary with underlying cardiac risk.[6]

Applicability Depends on Context and Cannot Be Assessed With a Universal Rating System

Several investigators have proposed series of questions or checklists for rating applicability.[5,7-9] Critical elements vary with the clinical area and intervention studied, thus it is

not clear that developing a single universal checklist is feasible. For example, there is little overlap between the items identified by Piboleau[9] for assessing applicability of orthopedic studies and those identified for assessing community interventions by Green.[8] Since we also found no empiric data validating the use of checklists for rating applicability across a range of clinical topics, we do not recommend use of any single checklist to rate applicability, but existing ones may provide a useful guide for factors to consider.

Applicability Is Best Reported Separately From the Strength of a Body of Evidence

GRADE incorporates considerations of applicability or directness into their assessments of the quality (or strength) of evidence from a body of studies, defined as the "level of confidence that an estimate of effect is correct."[4] This approach, however, does not recognize that a body of evidence with limited applicability may nonetheless provide strong evidence for one set of decisions or users but poor evidence for another. For example, early trials of thrombolysis for acute stroke may provide strong evidence for clinical decisions in specialized stroke centers but poor evidence for decisions in small rural emergency departments. We thus recommend reporting and discussing factors that limit or strengthen applicability of a body of evidence separately, rather than including it with judgments about risk of bias and other factors to determine overall quality or strength of evidence.[10] It may be reasonable to incorporate applicability into strength of evidence where reviews are created with a single primary audience in mind [11] with common, well-defined perspectives—for example, reviews for the U.S. Preventive Services Task Force incorporate into their recommendations considerations about whether the evidence is applicable to a representative North American population cared for in primary care.[12]

Four Specific Steps

We outline below four steps in assessing and reporting applicability. We distinguish the reporting and assessment of applicability of individual studies (steps 1-3) from reporting and assessment of the applicability of a body of evidence (step 4).

Step 1. Determine the Most Important Factors that May Affect Applicability

Identify potential factors. The PICOS is a useful way of organizing factors that may affect applicability. Including "setting" separately may capture information not reliably reported in population or intervention characteristics. For example, studies that recruit or treat patients in specialty settings may not be applicable to primary care populations due to differences that may not be apparent from other reported details.

Table 1 lists a variety of factors organized by the PICOS framework that may limit the applicability of individual research studies. Many of these elements are routinely captured in most systematic reviews (for example, demographics, event rates, etc.) but many other specific factors are often overlooked.

Table 1. Characteristics of individual studies that may affect applicability

	Condition that may limit applicability	Example	Feature that should be abstracted into evidence tables
Population	Narrow eligibility criteria and exclusion of those with comorbidities	**In the FIT trial,13 the trial randomized only 4000 of 54,000 originally screened. Participants were healthier, younger, thinner, and more adherent than typical women with osteoporosis**.	Eligibility criteria and proportion of screened patients enrolled; presence of comorbidities
	Large differences between demographics of study population and community patients	Cardiovascular clinical trials used to inform Medicare coverage enrolled patients who were significantly younger (60.1 vs. 74.7 years) and more likely to be male (75% vs. 42%) than Medicare patients with cardiovascular disease.[14]	Demographic characteristics: age, sex, race and ethnicity
	Narrow or unrepresentative severity, stage of illness, or comorbidities	Two-thirds of patients treated for congestive heart failure (CHF) would have been ineligible for major trials. Community patients had less severe CHF, more comorbidities and were more likely to have had a recent cardiac event or procedure.[14]	**Severity or stage of illness; comorbidities; referral or primary care population; volunteers vs. population-based recruitment strategies**.
	Run in period with high-exclusion rate for nonadherence or side effects	Trial of etanercept for juvenile arthritis used an active run in phase and excluded children who had side-effects, resulting in study with low rate of side-effects.[13]	Run in period; include attrition before randomization and reasons (nonadherence, side-effects, nonresponse)[14,15]
	Event rates much higher or lower than observed in population-based studies	In the Women's Health Initiative trial of post-menopausal hormone therapy, the relatively healthy volunteer participants had a lower rate of heart disease (by up to 50%) than expected for a similar population in the community.[16]	Event rates in treatment and control groups
Intervention	Doses or schedules not reflected in current practice	Duloxetine is usually prescribed at 40-60mg/d. Most published trials, however, used up to 120 mg/d.[17]	Dose, schedule, and duration of medication
	Intensity and delivery of behavioral interventions that may not be feasible for routine use	Studies of behavioral interventions to promote healthy diet employed high number and longer duration of visits than is available to most community patients.[18]	Hours, frequency, delivery mechanisms (group vs. individual) and duration.
	Monitoring practices or visit frequency not used in typical practice	Efficacy studies with strict pill counts and monitoring for antiretroviral treatment does not always translate to effectiveness in real world practice.[19]	Interventions to promote adherence (e.g., monitoring, frequent contact). Incentives given to study participants.
	Older versions of an intervention no longer in common use	Only one of 23 trials comparing coronary artery bypass surgery with percutaneous coronary angioplasty used the type of drug eluting stent that is currently used in practice.[15]	Specific product and features for rapidly changing technology
	Cointerventions that are likely to modify effectiveness of therapy	Supplementing zinc with iron reduces the effectiveness of iron alone on hemoglobin outcomes.[20] Recommendations for iron are based on studies examining iron alone, but patients most often take vitamins in a multivitamin form.	Cointerventions
	Highly selected intervention team or level of training/proficiency not widely available	Trials of carotid endarterectomy selected surgeons based on operative experience and low complication rates and are not representative of community experience of vascular surgeons.[21]	Selection process, training and skill of intervention team.

Table 1. Characteristics of individual studies that may affect applicability (continued)

	Condition That May Limit Applicability	Example	Feature that should be abstracted
Comparator	Inadequate dose of comparison therapy	A fixed dose study[20] by the makers of duloxetine compared 80 and 120 mg/d of duloxetine (high dose) with 20 mg of paroxetine (low dose).[22]	Dose and schedule of comparator, if applicable
	Use of substandard alternative therapy	In early trials of magnesium in acute myocardial infarction, standard of treatment did not include many current practices including thrombolysis and beta-blockade.[23]	Relative comparability to the treatment option.
Outcomes	Composite outcomes that mix outcomes of different significance	Cardiovascular trials frequently use composite outcomes that mix outcomes of varying importance to patients.[24]	Effects of intervention on most important benefits and harms, and how they are defined
	Short-term or surrogate outcomes	Trials of biologics for rheumatoid arthritis used radiographic progression rather than symptoms.[25] Trials of Alzheimer's disease drugs primarily looked at changes in scales of cognitive function over 6 months which may not reflect their ability to produce clinically important changes such as institutionalization rates.[26]	How outcome defined and at what time
Setting	Standards of care differ markedly from setting of interest	Studies conducted in China and Russia examined the effectiveness of self breast exams on reducing breast cancer mortality, but these countries do not routinely have concurrent mammogram screening as is available in the United States.[27]	Geographic setting
	Specialty population or level of care differs from that seen in community	Early studies of open surgical repair for abdominal aortic aneurysms found an inverse relationship between hospital volume and short-term mortality.[28]	Clinical setting (e.g. referral center vs. community)

Select a limited number of the most important factors that may affect applicability. Table 1 presents a wide range of items to consider. It is not feasible or necessary to record and report all of these items regardless of topic. Reviewers must instead exercise judgment to select a subset of the most important study parameters for the clinical topic. Foremost are any factors that have been associated with differences in treatment outcomes.

The observation that effectiveness of an intervention varies in different populations or settings is known as *heterogeneity of treatment effect.*[29] One cause of heterogeneity is true *effect modification*, defined when characteristics of the patient, intervention, or setting modify the relative effect of the intervention on the main outcome. Rothwell[30] notes the example where the benefits of carotid endarterectomy after a transient ischemic attack vary dramatically with the severity of the carotid stenosis and the timing of the surgery. We recommend reviewers solicit input from clinical experts and stakeholders to identify specific biologic, clinical, or health system factors that are known or suspected effect modifiers. Emphasis should be given to factors where statistically significant interactions or sub-group differences have been demonstrated in multiple studies. These factors should be identified a priori and stated in the protocol which factors will be captured in data extraction. For example, if age is a known effect modifier, evidence from studies of middle-aged adults will not be applicable to older populations. Additionally, emerging evidence has identified a number of genetic variations that modify the effectiveness of various drugs.

A more common source for heterogeneity in treatment effect is varying baseline rates of events. Even when an intervention has constant relative effects, *the absolute benefits and harms* will vary among populations with different baseline risks. For example, although statins reduce risks of fatal and nonfatal coronary events comparably in populations at high or lower risk of heart disease, the absolute benefits in high-risk populations such as those with a previous myocardial infarction are much larger (and thus not applicable) to lower risk populations.[31] Reviewers should routinely capture information on baseline or control group risk as a factor that may affect applicability.

Finally, intervention features may affect the *ability to generalize the effectiveness or safety of the intervention to use in everyday practice*. For example, outcome studies suggest that mortality after carotid surgery is affected by the experience of the center where surgery is performed, thus evidence from trials at selected tertiary centers may not be applicable to most community populations.[21] Clinical experts, population based surveys, outcome studies, and disease or procedure registries can provide information on current treatment context and whether typical populations, settings and interventions are represented in available studies.

Step 2. Systematically Abstract and Report Key Characteristics that May Affect Applicability in Evidence Tables; Highlight any Effectiveness Studies

Once the most important factors are selected, reviewers should abstract the relevant information into evidence tables under the relevant PICOS categories. Evidence tables should also highlight effectiveness trials. These studies (also referred to as "pragmatic" or "practical" trials) are designed to give more broadly applicable results than more common efficacy studies,[32] typically by enrolling more representative populations, letting interventions vary as they often do in practice, and focusing on the most important clinical benefits and harms.[32-34] Published criteria can be used to distinguish effectiveness trials from efficacy trials.[35,36] If data from both efficacy and effectiveness studies are available, comparing findings may indicate whether more narrowly

designed studies are applicable to broader populations. At the same time, reviewers must also examine whether effectiveness studies conceal important subgroup differences.[33]

Step 3. Make and Report Judgments About Major Limitations to Applicability of Individual Studies

Describe impact of applicability on interpretation of individual studies. To make applicability information useful, a review should address how specific aspects of the design of the study affected the final population or the quality of the intervention, and how greatly (and in which direction) these may differ from more representative populations in practice. For example, surgical studies that recruited surgeons based on good operative outcomes had significantly lower perioperative mortality than those observed in national Medicare hospitals,[21] (1.4 percent vs. 1.7, 1.9, or 2.5 percent for those high, average, or low volume). Thus, the balance of benefits and harms in the study are likely to overestimate those that would be expected for older patients treated in the community. Although this step involves judgment, such judgments can be made more explicit by considering how different this study is from a true *effectiveness* study and how those differences might have affected baseline risks of the population or the effectiveness or harms of the intervention.

Step 4. Consider and Summarize the Applicability of a Body of Evidence

Applicability of a body of studies is not the same as applicability of the individual studies. A collection of studies addressing one intervention or comparison generally provides more broadly applicable evidence than any individual study. Consistent results across studies that represent an array of different populations and settings increases our confidence that results are applicable across a broad set of conditions. For example, the individual trials of statin drugs to treat high cholesterol each selected specific and discrete populations, used different drugs, different dosages, and different cointerventions. While few would qualify as effectiveness trials individually, consistent findings across trials enrolling populations of differing risks, nationalities, and underlying conditions provides evidence that the benefits of statin drugs apply across a broad range of patients.

When the number of studies is large enough, the influence of specific factors (for example, age or gender) may be explored in additional analysis such as a subgroup analysis or meta-regression. If studies vary substantially in the underlying risk or event-rate, reviewers can test whether the effectiveness of treatment varies in high- and low-risk populations and judge which studies most closely approximate the typical risk in a more representative sample—this may require analysis of more representative registry or cohort data. We caution that meta-regression or other comparisons based on group level characteristics, such as the proportion of women in each trial, can be prone to bias (the "ecological fallacy").[37] Meta-analysis based on individual-patient data is more powerful.[37]

Describe the limitations of aggregate evidence using PICOS structure. Describe whether the collected body of evidence includes relevant populations, interventions, and appropriate comparisons, includes most important outcomes, and uses representative settings. Note whether studies share features that limit applicability—for example, did all the studies exclude older, sicker patients? Where studies vary in important features, inspect whether this variation is associated with differences in measures of effectiveness or safety. Reviewers should then describe how the available body of evidence differs from "ideal" evidence to answer the question

and indicate which characteristics of the evidence limit the applicability of the available evidence.

Use a summary table for applicability to highlight significant limitations to applicability. When there is a large body of evidence or when there are significant issues relevant to applicability, a summary table displays important applicability issues across a diverse body of evidence (see Table 2). One table may suffice for multiple questions if the same collection of studies is used to answer multiple questions (for example, the benefits and harms of an intervention). Critical concerns about applicability, however, can and should be described in the text.

Table 2. Elements to be included in a summary table characterizing the applicability of a body of studies

Domain	Description of applicability of evidence
Population	Describe general characteristics of enrolled populations, how this might differ from target population, and effects on baseline risk for benefits or harms. Where possible, describe the proportion with characteristics potentially affecting applicability (e.g. % over age 65) rather than the range or average.
Intervention	Describe general characteristics and range of interventions and how they compare to those in routine use and how this might affect benefits or harms from the intervention
Comparators	Describe comparators used. Describe whether they reflect best alternative treatment and how this may influence treatment effect size
Outcomes	Describe what outcomes are most frequently reported and over what time period. Describe whether the measured outcomes and timing reflect the most important clinical benefits and harms.
Setting	Describe geographic and clinical setting of studies. Describe whether or not they reflect the settings in which the intervention will be typically used and how this may influence the assessment of intervention effect.

Include the applicability of evidence in summary statements and tables addressing key questions. Comparative effectiveness reviews typically describe overall conclusions on the key questions in summary text and tables, including the effect for important outcomes and a characterization of the strength of evidence. Since we recommend separating applicability from "quality of evidence," summary conclusions should also describe the key issues affecting applicability. For example, when concluding that there is high quality evidence that carotid endarterectomy can reduce the risk of stroke and death in patients with asymptomatic carotid stenosis, it is important to specify that the evidence is applicable to patients treated at centers where the perioperative risk is less than 3 percent and who were followed an average of 4 years.[38]

Limitations of This Approach

This paper provides guidance for conducting comparative effectiveness reviews or other systematic reviews which address relatively broad clinical or policy questions in representative patient populations—for example, what is the comparative effectiveness of carotid endarterectomy vs. carotid stenting for patients with carotid stenosis? When the clinical question of interest has a much narrower focus—for example, is carotid stenting as safe and effective as carotid endarterectomy for women with a recent transient ischemic attack—it is better to restrict the review to studies which report results directly applicable to the specific question.

A related but distinct set of considerations are involved in applying evidence clinical decisions for an individual patient. Individual studies and systematic reviews give the best

estimates of the average effects but these averages may not apply to many individuals.[29] As Sackett has noted, clinical decisions need to incorporate best evidence, individual patient information (e.g. disease severity, life-expectancy, comorbidity), and individual preferences.[39]

Conclusions

Understanding the applicability of scientific evidence is an important but under-examined aspect of the systematic review process. Frequently, systematic reviews collect and present an abundance of details on elements of individual studies that are relevant to the applicability of the results, but few reviews organize this information to focus attention on specific concerns related to applicability. We describe an explicit approach to identifying, reporting and synthesizing information to allow consistent and transparent consideration of the applicability of the evidence in a systematic review. Although the exact process needs to be flexible and will likely evolve, attention to the general concepts described here will improve the ability of clinicians and policy makers to understand better to whom the conclusions of a systematic review apply, and under what conditions. In some instances it may lead to more cautious conclusions due to limitations in applicability. In others, a careful consideration of applicability may give decision makers greater confidence that the evidence summarized is appropriate and applicable for clinical and policy decisions. In both cases, it should improve the usefulness of systematic reviews, in informing practice and policy.

References

1. Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and the Congress on Comparative Effectiveness Research. Available at: http://www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf. Accessed June 30, 2009.
2. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2, updated September 2009. The Cochrane Collaboration 2009. Available at: http://www.cochrane-handbook.org.
3. Shadish, W, Cook T, Campbell D. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2002.
4. Guyatt GH, Oxman AD, Kunz R, et al. What is "quality of evidence" and why is it important to clinicians? BMJ 2008 May 3;336(7651):995-998.
5. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" Lancet 2005 Jan 1-7;365(9453):82-93.
6. Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the Effective Health-Care Program. J Clin Epidemiol 2010 May;63(5):502-512.
7. Bornhöft G, Maxion-Bergemann S, Wolf U, et al. Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. BMC Med Res Methodol 2006 Dec 11;6:56
8. Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof 2006 Mar; 29(1):126-153
9. Pibouleau L, Boutron I, Reeves BC, et al. Applicability and generalisability of published results of randomised controlled trials and non-randomised studies evaluating four orthopaedic procedures: methodological systematic review. BMJ 2009 Nov 17;339:b4538.
10. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions. Agency for Healthcare Research and Quality and the Effective Health-Care Program. J Clin Epidemiol 2010 May;63(5):513-523.

11. Falck-Ytter Y, Schünemann H, Guyatt G. AHRQ series commentary 1: rating the evidence in comparative effectiveness reviews. J Clin Epidemiol 2010 May;63(5):474-475.

12. Guirguis-Blake J, Calonge N, Miller T, et al. Current processes of the U.S. Preventive Services Task Force: refining evidence-based recommendation development. Ann Intern Med 2007 Jul 17;147(2):117-122.

13. Cummings SR, Black DM, Thompson DE, et al. Effect of alendronate on risk of fracture in women with low bone density but without vertebral fractures: results from the fracture intervention trial. JAMA 1998;280(24):2077-2082

14. Dhruva SS, Redberg RF. Variations between clinical trial participants and Medicare beneficiaries in evidence used for Medicare National Coverage Decisions. Arch Intern Med 2008 Jan; 169(2):136-140

15. Bravata DM, McDonald KM, Gienger AL, et al. Comparative Effectiveness of Percutaneous Coronary Interventions and Coronary Artery Bypass Grafting for Coronary Artery Disease. Comparative Effectiveness Review No. 9. (Prepared by Stanford-UCSF Evidence-based Practice Center under Contract No. 290-02-0017.) Rockville, MD: Agency for Healthcare Research and Quality; October 2007.

16. Anderson GL, Limacher M, Assaf AR, et al. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. JAMA 2004 Apr 14;291(14):1701-1712.

17. Gartlehner G, Hansen RA, Thieda P, et al. Comparative Effectiveness of Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression. Comparative Effectiveness Review No. 7. (Prepared by RTI International-University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality; January 2007.

18. Whitlock EP, O'Connor EA, Williams SB, et al. Effectiveness of Weight Management Programs in Children and Adolescents. Evidence Report/Technology Assessment No. 170 (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024). AHRQ Publication No. 08-E014. Rockville, MD: Agency for Healthcare Research and Quality; September 2008.

19. Fletcher CV. Translating efficacy into effectiveness in antiretroviral therapy: beyond the pill count. Drugs 2007;67(14):1969-1979.

20. Walker, CF, Kordas K, Stoltzfus, RJ, et al. Interactive effects of iron and zinc on biochemical and functional outcomes in supplementation trials. Am J Clin Nutr 2005 82: 5-12.

21. Wennberg D, Lucas F, Birkmeyer J, et al. Variation in carotid endarterectomy mortality in the Medicare population. JAMA 1998;279:1278-1281.

22. Detke MJ, Wiltse CG, Mallinckrodt CH, et al. Duloxetine in the acute and long-term treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. Eur Neuropsychopharmacol 2004 Dec;14(6):457-470.

23. Li J, Zhang Q, ZhangM, et al. Intravenous magnesium for acute myocardial infarction. Cochrane Database of Systematic Reviews 2007, Issue 2. Art. No.: CD002755. DOI: 10.1002/14651858.CD002755.pub2.

24. Ferreira-González I, Permanyer-Miralda G, Domingo-Salvany A, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 2007;334;786; originally published online 2 Apr 2007

25. Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. J Clin Epidemiol 1997 Oct;50(10):1089-1098.

26. Hansen RA, Gartlehner G, Kaufer D, et al. Drug class review of Alzheimer's drugs. Final report. 2006. Available at: http://www.ohsu.edu/drugeffectiveness/reports/final.cfm.

27. Humphrey L, Chan BKS, Detlefsen S, et al. Screening for Breast Cancer. Prepared by Oregon Health Sciences University under Contract No. 290-97-0018. Rockville, MD. Agency for Healthcare Research and Quality; August 2002.

28. Wilt TJ, Lederle FA, MacDonald R, et al. Comparison of Endovascular and Open Surgical Repairs for Abdominal Aortic Aneurysm. Evidence Report/Technology Assessment No. 144. (Prepared by the University of Minnesota Evidence-based Practice Center under Contract No. 290-02-0009.) AHRQ Publication No. 06-E017. Rockville, MD: Agency for Healthcare Research and Quality; August 2006.

29. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004;82(4):661-687.

30. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. PLoS Clin Trials 2006 May;1(1):e9

31. National Institute for Health and Clinical Excellence. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. London: NICE; 2008. Available at: www.nice.org.uk/ CG67

32. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003 Sep 24;290(12):1624-1632.

33. Godwin M, Ruhland L, Casson I, et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003 Dec 22;3:28.

34. Atkins D. Creating and synthesizing evidence with decision makers in mind: integrating evidence from clinical trials and other study designs. Med Care 2007 Oct; 45(10 Supl 2):S16-S22.

35. Gartlehner G, Hansen RA, Nissman D, et al. A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006 Oct;59(10):1040-1048. Epub 2006 Aug 4.

36. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009 May;62(5):464-475.

37. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010 Feb 5;340:c221. doi: 10.1136/bmj.c221.

38. Chambers BR, Donnan G. Carotid endarterectomy for asymptomatic carotid stenosis. Cochrane Database of Systematic Reviews 2005, Issue 4. Art. No.: CD001923. DOI: 10.1002/14651858.CD001923.pub2.

39. Sackett DL, Richardson WS, Rosenberg W, et al. Evidence-based medicine—how to practice and teach EBM. New York: Churchill Livingstone; 1997.

Appendix A. Example Adapted from Comparative Effectiveness Review of Therapies for Clinically Localized Prostate Cancer[A1]

We have augmented consideration of applicability from a previous comparative effectiveness review[A1] illustrating the different steps for assessing and reporting the applicability of the evidence to the following question:

How do the benefits and harms of radical prostatectomy compare to watchful waiting for treatment of early organ-confined prostate cancer?

Step 1. Determine the Most Important Factors that May Affect Applicability

In order to determine the important factors, the reviewers must consider the underlying biology and epidemiology as well as the historical and current clinical practice context.

Epidemiologic studies indicate that prostate cancer prognosis is tied to *grade* and, to a lesser extent, *stage* of cancer. Cancer registries in the United States indicate that most localized cancers are detected by PSA testing (Stage T1c), with the majority diagnosed in men over age 65. Clinical experts think that *age and comorbidity* affect benefits and risks of aggressive therapy (by creating competing risks which reduce the benefits of aggressive interventions and by increasing risks of surgery). Specific *cointerventions* or *surgical techniques* (e.g. nerve-sparing approaches or adjuvant hormonal therapy) and *experience of the participating centers and surgeons* may influence both the effectiveness of treatment and adverse event rates.

Step 2. Systematically Abstract and Report Characteristics that May Affect Applicability in Evidence Tables; Highlight Any Effectiveness Studies

Table A-1 is an abbreviated version of an evidence table, into which the reviewer extracts relevant data from individual studies, used to judge both internal validity and applicability. However, this example table focuses only on data related to applicability of the study.

Table A-1. Example evidence table of individual studies with key applicability factors abstracted and judgment of applicability

Trial (including date, setting)	Population Demographic, Disease state	Intervention	Comparator	Outcomes and timing	Comments
Bill-Axelson et al.[A2] (SPCG-4) 1989-1999, Sweden	Mean age 65 78% T2 60% Gleason 6 or lower. Few detected by PSA	Radical prostatectomy at 18 centers; standard current protocol	Watchful waiting with deferred hormonal therapy	Prostate-specific antigen and all cause mortality; metastasis and disease progression; median follow-up of 8.3 years	Some indications of an effectiveness trial. Unclear how highly selected the enrolled patients were. Limited standardization of the intervention. Unclear whether the participating centers and surgeons are representative of the larger population.
Iversen et al.[A3] 1967-1975 Denmark	Mean age 64.2 46.5% Stage 2 86.5% Gleason 6 or lower. None detected by PSA.	Radical prostatectomy in one Veterans Administration center, protocol from 1967-1975	Watchful waiting with oral placebo	Overall mortality; Median follow-up 23 years	Results may not be applicable to current practices due to the evolving techniques in both stage and grade classification since PSA screening.

Step 3. Make and Report Judgments About Major Limitations to Applicability of Individual Studies

Once the appropriate data for assessing applicability of individual studies has been identified, the reviewer must then consider what impact it will have when interpreting the results of the study in relation to the question being asked.

The reviewer can then highlight and summarize the key concerns or strengths of an individual study for its applicability to the question, highlighting effectiveness studies. We illustrate how this might be done in the comments column of Table A-1 above.

Step 4. Consider and Summarize the Applicability of a Body of Studies

After identifying the major strengths and limitations in applicability for individual studies, the reviewer must then consider the applicability of the body of evidence and considering how the limitations may impact the interpretation of the evidence in answering the question. In order to do this, it may be helpful to use a summary table for applicability, as illustrated in Table A-2.

Table A-2. Example summary table characterizing the applicability of a body of studies

Domain	Description of applicability of evidence
Population	Available trials included few patients with PSA detected by screening (T1c), whose prognosis may be different. The age of enrolled patients was representative of prostate cancer patients in the community, but subgroup results from one study suggest that benefits of treatment may be smaller in patients over age 65 than those under age 65.
Intervention	The prostatectomy treatment in the Scandinavian study[2] is applicable to current surgical methods although it is not clear if nerve-sparing surgery was common. The smaller trial[3] was conducted over 20 years ago and may not be applicable.
Comparators	Watchful waiting is an appropriate comparator in both studies but only the more recent study used hormonal therapy for patients whose disease progresses.
Outcomes	Available trials use a reasonable array of health outcomes. Additional follow-up from one study suggests that outcomes at 10 years are representative of longer-term outcomes. For older patients, prostate cancer mortality may represent a small portion of overall mortality and thus be less relevant than overall mortality.
Setting	One study was conducted across a broad cross section of Scandinavian centers, whereas the other was conducted in a highly selected population from one Danish Veterans Administration center in the 1960s-1970s. It is not clear in what direction this may affect the results. They may be a healthier population from having regular access to medical care, but may be more likely to have other comorbidities such as heart disease than a highly selected population.

With use of a summary applicability table, it becomes easier for a reviewer to describe in the text how aspects of the study may impact the interpretation of the study results in answering the question. An example of a text summary of applicability and their implications is provided below.

Two trials have addressed the benefits of surgical therapy compared to deferred therapy or watchful waiting. Results are dominated by one trial, which demonstrated important but modest benefits of prostatectomy. There are important concerns about the applicability of this evidence to the population of interest. These results are most applicable to patients under 65 with T2 prostate cancer but cannot be assumed to apply to the largest group of prostate cancer patients in the United States, those with cancers detected by PSA screening (T1c). Such patients have a substantially better untreated prognosis and would be unlikely to benefit as much from surgery, at least over the 8 to 10 year time period of the available trials. Whether results apply to older patients is unclear. Patients over age 65 had smaller benefits in a subgroup analysis of the Swedish trial but this difference was not statistically significant; nonetheless the high risk of competing causes of death reduces the number of patients that will live long enough to benefit.

Finally, at the level of synthesis, the reviewer should describe the applicability of the evidence in the highest level of summary conclusions. This is often presented in the form of the summary table (Table A-3).

Table A-3. Example summary table for body of evidence

Comparison	Strength of Evidence	Conclusions with description of applicability
Radical prostatectomy vs. watchful waiting	Medium	Compared with men who used watchful waiting, men with localized prostate cancer detected by methods other than PSA testing and treated with radical prostatectomy (RP) experienced fewer deaths from prostate cancer and fewer distant metastases. The benefits of RP on cancer-specific and overall mortality appears to be limited to men under 65 years of age but is not dependent on baseline PSA level or histologic grade.

References

A1. Wilt TJ, Shamliyan T, Taylor B, et al. Comparative Effectiveness of Therapies for Clinically Localized Prostate Cancer. Comparative Effectiveness Review No. 13. (Prepared by Minnesota Evidence-based Practice Center under Contract No. 290-02-0009.) Rockville, MD: Agency for Healthcare Research and Quality; February 2008.

A2. Bill-Axelson A, Holmberg L, Ruutu M, et al. Scandinavian Prostate Cancer Screening Group Study No. 4. Radical prostatectomy versus watchful waiting in early prostate cancer. N Engl J Med 2005 May 12;12(19):1977-1984.

A3. Iversen P, Madsen PO, Corle DK. Radical prostatectomy versus expectant treatment for early carcinoma of the prostate. Twenty-three year followup of a prospective randomized study. Scan J Urol Nephrol Suppl 1995;172:65-72.