Current Methods of the U.S. Preventive Services Task Force

A Review of the Process

Russell P. Harris, MD, MPH, Mark Helfand, MD, MS, Steven H. Woolf, MD, MPH, Kathleen N. Lohr, PhD, Cynthia D. Mulrow, MD, MSc, Steven M. Teutsch, MD, MPH, David Atkins, MD, MPH
for the Methods Work Group, Third U.S. Preventive Services Task Force

Abstract: The U.S. Preventive Services Task Force (USPSTF/Task Force) represents one of several efforts to take a more evidence-based approach to the development of clinical practice guidelines. As methods have matured for assembling and reviewing evidence and for translating evidence into guidelines, so too have the methods of the USPSTF. This paper summarizes the current methods of the third USPSTF, supported by the Agency for Healthcare Research and Quality (AHRQ) and two of the AHRQ Evidence-based Practice Centers (EPCs).

The Task Force limits the topics it reviews to those conditions that cause a large burden of suffering to society and that also have available a potentially effective preventive service. It focuses its reviews on the questions and evidence most critical to making a recommendation. It uses analytic frameworks to specify the linkages and key questions connecting the preventive service with health outcomes. These linkages, together with explicit inclusion criteria, guide the literature searches for admissible evidence.

Once assembled, admissible evidence is reviewed at three strata: (1) the individual study, (2) the body of evidence concerning a single linkage in the analytic framework, and (3) the body of evidence concerning the entire preventive service. For each stratum, the Task Force uses explicit criteria as general guidelines to assign one of three grades of evidence: good, fair, or poor. Good or fair quality evidence for the entire preventive service must include studies of sufficient design and quality to provide an unbroken chain of evidence-supported linkages, generalizable to the general primary care population, that connect the preventive service with health outcomes. Poor evidence contains a formidable break in the evidence chain such that the connection between the preventive service and health outcomes is uncertain.

For services supported by overall good or fair evidence, the Task Force uses outcomes tables to help categorize the magnitude of benefits, harms, and net benefit from implementation of the preventive service into one of four categories: substantial, moderate, small, or zero/negative.

The Task Force uses its assessment of the evidence and magnitude of net benefit to make a recommendation, coded as a letter: from A (strongly recommended) to D (recommend against). It gives an I recommendation in situations in which the evidence is insufficient to determine net benefit.

The third Task Force and the EPCs will continue to examine a variety of methodologic issues and document work group progress in future communications.

Medical Subject Headings (MeSH): MEDLINE, preventive health services, evidence-based medicines, methods, practice guidelines (Am J Prev Med 2001;20(3S):21–35)

From the School of Medicine and Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill (Harris), North Carolina; Division of Medical Informatics and Outcomes Research, and Evidence-based Practice Center, Oregon Health Sciences University and Portland Veterans Affairs Medical Center (Helfand), Portland, Oregon; Department of Family Practice, Medical College of Virginia, Virginia Commonwealth University (Woolf), Fairfax, Virginia; Research Triangle Institute, Research Triangle Park, and University of North Carolina at Chapel Hill, Program on Health Outcomes and School of Public Health (Lohr), North Carolina; Department of Medicine, University of Texas Health Science Center (Mulrow), San Antonio, Texas; Outcomes Research and Management, Merck & Co., Inc. (Teutsch), West Point, Pennsylvania; Center for Practice and Technology Assessment, Agency for Healthcare Research and Quality (Atkins), Rockville, Maryland

Other members of the Methods Work Group include: Alfred O. Berg, MD, MPH, University of Washington School of Medicine; Karen B. Eden, PhD, Oregon Health Sciences University; John Feightner, MD, MSc, FCFP, University of Western Ontario–Parkwood

0749-3797/01/$–see front matter
PII S0749-3797(01)00261-6

Introduction

The U.S. Preventive Services Task Force (Task Force/USPSTF) represents one of several efforts by governments and national organizations to take a more evidence-based approach to the development of clinical practice guidelines. Guidelines developed by an evidence-based approach tend to be based on conclusions supported more by scientific evidence than by expert opinion.[1] Efforts are made to link the strength of recommendations to the quality of evidence; to make that linkage transparent and explicit, and to ensure that the review of evidence is comprehensive, objective, and attentive to quality.[2]

Methods for reviewing the evidence have matured over the years as groups have gained experience in developing evidence-based guidelines. Systematic searches of multiple bibliographic research databases help ensure thorough and unbiased identification of the relevant literature. Predetermined selection criteria minimize bias and improve the efficiency of reviewing that literature. Quality criteria developed by methodologists guide judgments of weaknesses and strengths of individual research studies. Frameworks and models explicitly define methods for rating and integrating multiple pieces of heterogeneous evidence.[3]

Methods for linking evidence and recommendations have also matured.[4] Initially the recommendations of the USPSTF and other evidence-based groups were strongly correlated with the research design of the most important studies. An A recommendation, for example, usually meant that use of the preventive service was supported by a randomized controlled trial (RCT).[5,6] Guideline developers now understand the need to consider the evidence as a whole, including the trade-offs among benefits, harms, and costs and the net benefit relative to other health care needs for optimal resource allocation.[7]

In the case of prevention, moreover, special scientific and policy considerations apply in reviewing evidence and setting policy. Preventive services require a distinctive logic in considering, for example, the incremental benefit of early detection or the ability of counselors to motivate behavior change. Because the populations affected by preventive care recommendations are often large and have no recognized symptoms or signs of the target condition, harms incurred by even a small percentage can affect a large number of people. Thus, the potential for doing greater harm than good must be taken seriously.

In the context of these methodologic advances and with an awareness of the many unresolved issues for which sound methods are lacking, the third Task Force formed a methods subcommittee (Methods Work Group). It comprises members of the Task Force, representatives of the Canadian Task Force on Preventive Health Care, staff of the two Evidence-based Practice Centers (EPCs) that support the Task Force, and staff of the Agency for Healthcare Research and Quality (AHRQ). The mission of the Work Group is to revisit methods used by previous U.S. Preventive Services Task Forces, to develop more sophisticated methods to be used in current work, and to understand better the theoretical considerations for problems that lack easy answers.

The discussions of this group and subsequent discussions by the entire Task Force have led to several modifications of Task Force methods and identified areas that need further examination. This article describes the methods in current use by the third USPSTF. As the Task Force identifies better ways to do its work, the Methods Work Group will explore additional revisions and refinements to its methods.

We discuss these changes in the sequence of steps of recommendation development: scope and selection of topics, review of the evidence, assessing the magnitude of net benefit, extrapolation and generalization, translating evidence into recommendations, drafting the report, and external review.

Scope and Selection of Topics

Scope

In defining its scope of interest, the Task Force must consider types of services, populations of patients and providers, and sites for which its recommendations are intended. Clarifying these definitions has both methodologic and practical importance. Resource limitations make it impossible for the Task Force to review evidence for all services that prevent disease; the project must, therefore, set boundaries.

The third Task Force has retained the previous policy of focusing on screening tests, counseling interventions, immunizations, and chemoprevention delivered to persons without recognized symptoms or signs of the target condition.

As in the past, this Task Force decided not to make recommendations concerning services to prevent complications in patients with established disease (e.g., coronary artery disease and diabetes). It does, however, make recommendations for preventing morbidity or

Hospital; Susan Mahon, MPH, Oregon Health Sciences University; and Michael Pignone, MD, MPH, University of North Carolina School of Medicine.

Address correspondence to: Russell P. Harris, MD, MPH, Cecil G. Sheps Center for Health Services Research, CB# 7590, 725 Airport Rd., The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7590. E-mail: rharris@med.unc.edu.

Reprints are available from the AHRQ Web site at www.ahrq.gov/clinic/uspstfix.htm, through the National Guideline Clearinghouse (www.guideline.gov), or in print through the AHRQ Publications Clearinghouse (1-800-358-9295).

mortality from a second condition among those who have a different established disease.

The Task Force does make recommendations for people at different levels of risk for a condition. Many people in the general population have one or more risk factors for the Task Force's target conditions. Because the balance between benefits and harms sometimes differs between people at higher risk and those at lower risk, Task Force recommendations may vary across these different groups.

Although the Task Force does not conduct systematic searches of evidence for services to prevent complications in people with established disease, it may cite such studies when they are relevant for people without established disease. Often, compelling evidence that screening tests and treatments can reduce morbidity and mortality comes from patients with extant disease rather than from asymptomatic populations. For example, the review of lipid screening in this supplement [8] would be incomplete if it did not discuss studies of the efficacy of statins in patients with coronary artery disease.

The populations for whom Task Force recommendations are intended include patients seen in traditional primary care or other clinical settings (e.g., dieticians' offices, cardiologists' offices, emergency departments, hospitals, school-based clinics, urgent care facilities, student health clinics, family planning clinics, nursing homes, and homes). As before, the third Task Force has excluded consideration of preventive services outside the clinical setting (e.g., nonclinic-based programs at schools, worksites, and shopping centers), reserving this analysis to the work of the Centers for Disease Control and Prevention's (CDC) *Guide to Community Preventive Services* [9] effort. For selected topics, however, the Task Force may examine evidence from community-based settings to evaluate the effectiveness of interventions conducted in the clinical arena.

Selection of Topics

In the second edition of the *Guide to Clinical Preventive Services*,[6] the Task Force reviewed 70 preventive care topics, including more than 100 actual services. These had been selected on the basis of the burden of suffering to society or individuals and the potential effectiveness of one or more preventive interventions. The Task Force briefly considered using an explicit grading process for ranking the priority of topics, an exercise that was undertaken by the second Task Force with disappointing results, and for this reason the current Task Force did not pursue it.

Instead, the third Task Force started with the topics reviewed in the second *Guide to Clinical Preventive Services*.[6] From the 70 topics, the EPCs, AHRQ, and Task Force leaders identified 55 likely to have new evidence or continued controversy. For these 55 topics, the EPCs undertook limited literature searches and prepared brief summaries of the new evidence, current controversies, and critical issues. The EPCs prepared similar summaries of 15 new topics suggested by previous Task Force members, the public, outside experts, federal agencies, and health care organizations. AHRQ and the EPCs also invited about 60 private health and consumer groups and federal agencies to rate the need to update old chapters and to nominate new topics.

Based on this information, the USPSTF ranked the priority of topics at its first meeting in November 1998. It initially assigned 12 topics to the two EPCs (six to each EPC) for review and has subsequently added more topics in a phased schedule (Table 1).

The responsible EPC assigns a lead author and a variable number of additional local personnel to each topic. The Task Force assigns two or three of its own members ("Task Force liaisons") to collaborate on the review. The local EPC group and the Task Force liaisons constitute the "topic team" for each review. The EPCs make certain that all topic team personnel are trained in Task Force methods and the content area of the review.

Review of the Evidence

Intensity

Current methods for conducting systematic reviews emphasize a comprehensive literature search and evaluation and detailed documentation of methods and findings.[10] An advantage of this approach is that it avoids the tendency of some guideline panels to cite evidence selectively in support of their recommendations. This approach also enables others outside the process to understand, judge, and replicate the interpretation of the evidence. The disadvantage of this approach is that it produces long, detailed reports of interest to a minority of readers and of limited value to busy clinicians. The process is also resource intensive and requires months of work and considerable expenditures for literature searches and staff. Despite the disadvantages, many evidence-based groups use this approach when reviewing evidence.

For a group such as the Task Force and its EPCs, which must examine multiple topics at once, limited resources and time require compromises in the intensity of reviews. Full-scale systematic reviews for every topic considered are not possible. One strategy for striking a balance, already noted, is topic prioritization. Another strategy, initiated by the second Task Force, is to focus the review on the questions and evidence most critical to making a recommendation.

Setting the Focus for Admissible Evidence

Analytic framework. The second Task Force introduced diagrams, called "causal pathways," to map out

Table 1. Topics completed or under review by the third U.S. Preventive Services Task Force

Evidence-based Practice Center	
Research Triangle Institute–University of North Carolina	**Oregon Health Sciences University**
Updates	**Updates**
Screening for and treating adults for lipid disorders	Screening for breast cancer
Screening for type 2 diabetes mellitus	Screening for skin cancer
Counseling in the clinical setting to prevent unintended pregnancy	Counseling to prevent skin cancer
Counseling to promote a healthy diet	Screening for family violence
Screening for visual impairment in children aged 0 to 5 years	Screening for problem drinking
Screening for depression	Counseling to prevent youth violence
Screening for cervical cancer	Postmenopausal hormone chemoprevention
Screening for prostate cancer	Screening for chlamydial infection
Screening for colorectal cancer	Universal newborn hearing screening
Aspirin chemoprevention for the primary prevention of cardiovascular events	Screening for lung cancer
Screening for hypertension	Screening for ovarian cancer
Screening for gestational diabetes	Screening for iron deficiency anemia
Screening for asymptomatic coronary artery disease	Screening for neural tube defects
Screening for dementia	Screening for asymptomatic carotid artery stenosis
Screening for obesity	Screening for Down syndrome
Screening for suicide risk	Screening for osteoporosis
Counseling to prevent dental and periodontal disease	Counseling to promote physical activity
New	**New**
Chemoprevention of breast cancer	Screening for bacterial vaginosis in pregnancy
Screening for developmental delay	Counseling to promote breastfeeding
	Vitamin supplementation to prevent cancer and cardiovascular disease

the specific linkages in the evidence that must be present for a preventive service to be considered effective. The third Task Force retained these diagrams, renaming them "analytic frameworks." The analytic framework (Figures 1 and 2) uses a graphical format to make explicit the populations, preventive services, diagnostic or therapeutic interventions, and intermediate and health outcomes to be considered in the review. It demonstrates the chain of logic that evidence must support to link the preventive service to improved health outcomes.[11–13]

In the analytic framework, the arrows ("linkages"), labeled with a preventive service or a treatment, represent the questions that evidence must answer; dotted lines represent associations; rectangles represent the intermediate outcomes (rounded corners) or the health states (square corners) by which those linkages are measured. Figure 1 illustrates the analytic framework for a screening service, in which a population at risk (left side of the figure) undergoes a screening test to identify early-stage disease. A generic analytic framework for a counseling topic is given in Figure 2.

In Figure 1, an "overarching" linkage (arrow 1) above the primary framework represents evidence that directly links screening to changes in health outcomes. For example, an RCT of chlamydia screening established a direct, causal connection between screening and reduction in a pelvic inflammatory disease.[14] That is, a single body of evidence establishes the connection between the preventive service (screening) and health outcomes.

When direct evidence is lacking or is of insufficient quality to be convincing, the Task Force relies on a chain of linkages to assess the effectiveness of a service. In Figure 1, these linkages correspond to key questions about the accuracy of screening tests (arrow 3), the efficacy of treatment (arrow 4 or arrow 5 for intermediate or health outcomes, respectively), and the association between intermediate measures and health outcomes (dotted line 6). Intermediate outcomes (e.g., changes in serum lipid levels or eradication of chlamydia infection as measured by a DNA probe) are often used in studies as indicators of efficacy; health outcomes are measures that a patient can feel or experience, including death, quality of life, pain, and function. Curved arrows below the primary framework (arrows 7 and 8 in Figure 1) indicate adverse events or harms (ovals). Each arrow in the analytic framework relates to one or more "key questions" that specify the evidence required to establish the linkage (see the legends for Figures 1 and 2). These questions help organize the literature searches, the results of the review, and the writing of reports.

As can be seen in Figures 1 and 2, the framework supporting a service is considered indirect if two or more bodies of evidence are required to assess the effectiveness of the service. For example, no controlled studies provide direct evidence that screening for skin cancer lowers mortality.[15] To infer benefit, one must piece together evidence about the accuracy of the screening test, how much earlier screening detects skin cancer or its precursors than would be the case without

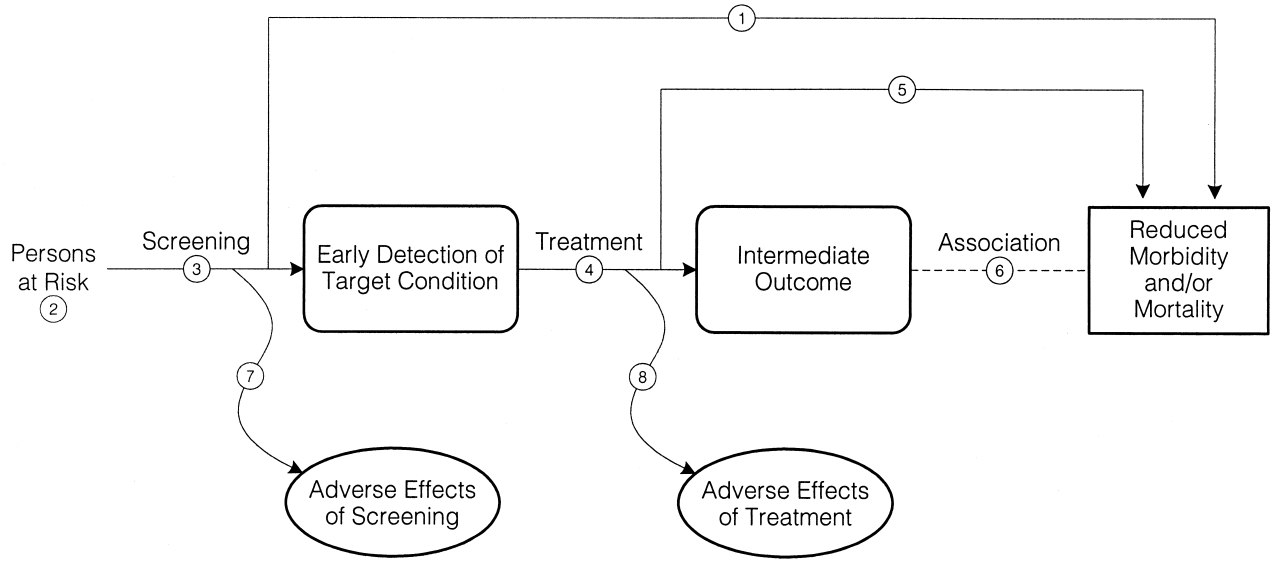

Figure 1. Generic analytic framework for screening topics. Numbers refer to key questions as follows: (1) Is there direct evidence that screening reduces morbidity and/or mortality? (2) What is the prevalence of disease in the target group? Can a high-risk group be reliably identified? (3) Can the screening test accurately detect the target condition? (a) What are the sensitivity and specificity of the test? (b) Is there significant variation between examiners in how the test is performed? (c) In actual screening programs, how much earlier are patients identified and treated? (4) Does treatment reduce the incidence of the intermediate outcome? (a) Does treatment work under ideal, clinical trial conditions? (b) How do the efficacy and effectiveness of treatments compare in community settings? (5) Does treatment improve health outcomes for people diagnosed clinically? (a) How similar are people diagnosed clinically to those diagnosed by screening? (b) Are there reasons to expect people diagnosed by screening to have even better health outcomes than those diagnosed clinically? (6) Is the intermediate outcome reliably associated with reduced morbidity and/or mortality? (7) Does screening result in adverse effects? (a) Is the test acceptable to patients? (b) What are the potential harms, and how often do they occur? (8) Does treatment result in adverse effects?

screening, the existence of effective treatment, whether treatment at an earlier stage improves health outcomes, and the existence and magnitude of associated harms. These criteria are similar to those outlined by the World Health Organization[16] and by Frame and Carlson.[17]

Admissible evidence. The third Task Force focuses its reviews primarily on the evidence most likely to influence recommendations. For example, it maintains the tradition of giving greater weight to evidence that preventive services influence health outcomes rather than intermediate outcomes. Although some intermediate outcomes (e.g., advanced-stage breast or colon cancer) are so closely associated with health outcomes that they are logical surrogates, many others (e.g., physiological changes or histopathologic findings) are less convincing because their reliability in predicting adverse health outcomes has weaker scientific support.[18,19] Accordingly, the topic teams often do not fully review studies that do not address outcomes of interest.

The topic team determines the bibliographic databases to be searched and the specific inclusion and exclusion criteria (i.e., admissible evidence) for the literature on each key question. Such criteria typically include study design, population studied, year of study, outcomes assessed, and length of follow-up. Topic teams specify criteria on a topic-by-topic basis rather than adhering to generic criteria. If high-quality evidence is available, the topic teams may exclude lower-quality studies. Conversely, if higher-quality evidence is lacking, the teams may examine lower-quality evidence. In general, the topic teams exclude non-English language references.

The second Task Force reviewed studies published through 1995. Thus, literature searches to update these topics usually extend from 1994 to the present, although new or refocused key questions may extend the search to older literature. For new topics, all searches begin with 1966 unless topic-specific reasons limit the search to a shorter time span or require an examination of even older literature. If a search finds a well-performed systematic review that directly addresses the literature on a key question through a given date, the topic team may use this review to capture the literature for those dates. The team can then restrict its own search to dates not covered by the existing systematic review.

The topic team documents these strategies for sharpening focus—the analytic framework, key questions, and criteria for admissible evidence—in an initial work plan. This work plan is presented to the Task Force at its first meeting after the topic has been assigned, allowing the Task Force the opportunity to modify the direction and scope of the review, as needed.

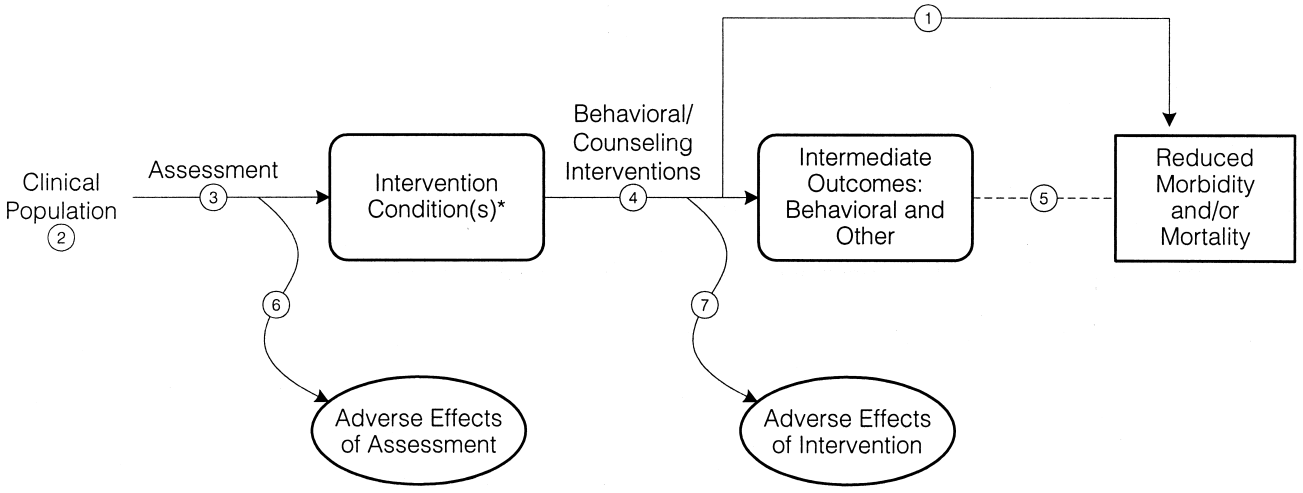

Figure 2. Generic analytic framework for counseling interventions. Numbers refer to key questions as follows: (1) Is there direct evidence that behavioral/counseling interventions reduce disease morbidity and/or mortality? (2) What is the prevalence of risky behavior(s) in the target group? Are there distinct patient groups for whom different intervention strategies apply? (3) Are there effective, feasible, and reliable assessment tools to identify those in need of interventions? (4) Does the behavioral/counseling intervention result in change in intermediate behavioral or other outcomes? (a) What are the essential elements of efficacious interventions? (b) Are there differences in efficacy in important patient subgroups? (c) How do intervention efficacy and effectiveness compare? (5) Does the behavior change lead to reduced morbidity and/or mortality? Do other intermediate outcomes related to the behavior change lead to decreased morbidity and/or mortality? (6) Is assessment for the behavioral/counseling intervention acceptable to patients? Does it result in adverse effects? (7) Is the behavioral/counseling intervention acceptable to patients? Does it result in adverse effects?
*An intervention condition is a distinct group identified through the assessment process that receives a different intervention. Evidence for each of the intervention conditions may be reviewed separately.

Literature Search and Abstraction

All searches involve at least the MEDLINE English-language database and the Cochrane Collaboration Library, using appropriate search terms to retrieve studies that meet the previously established inclusion and exclusion criteria. The search also includes other databases when indicated by the topic. The topic teams supplement these searches with references from reviews, current articles, and suggestions from experts in the field. Two members of the topic team (typically EPC staff) review abstracts of all articles. If either reviewer believes that the abstract meets the inclusion criteria, the EPC retrieves the full text of the article. The eligibility criteria are reapplied by one reviewer who, if the article is included, abstracts information about the patient population, study design, interventions (where appropriate), quality indicators, and findings.

Evaluating evidence: rethinking quality. The Methods Work Group, recognizing the central role that evaluating the quality of the evidence plays in the process of making evidence-based guidelines, focused much effort on this issue and decided to refine the process used by the previous Task Force. Specifically, the third Task Force adopted three important changes to the process: adding a rating of internal validity to the study design criterion for judging individual studies, explicitly assessing evidence at three different strata, and separating the magnitude of effect from the assessment of quality.

Evaluating quality at three strata: Stratum 1, the individual study. For some years, the standard approach to evaluating the quality of individual studies was based on a hierarchical grading system of research design in which RCTs received the highest score (Table 2). The maturation of critical appraisal techniques has drawn attention to the limitations of this approach, which gives inadequate consideration to how well the study was conducted, a dimension known as internal validity.[20] A well-designed cohort study may be more compelling than an inadequately powered or poorly conducted RCT.[21,22]

Table 2. Hierarchy of research design

I	Evidence obtained from at least one properly randomized controlled trial.
II–1	Evidence obtained from well-designed controlled trials without randomization.
II–2	Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group.
II–3	Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence.
III	Opinions of respected authorities, based on clinical experience, descriptive studies and case reports, or reports of expert committees.

Table 3. Criteria for grading the internal validity of individual studies

Study design	Criteria
Systematic reviews	• Comprehensiveness of sources/search strategy used • Standard appraisal of included studies • Validity of conclusions • Recency and relevance
Case–control studies	• Accurate ascertainment of cases • Nonbiased selection of cases/controls with exclusion criteria applied equally to both • Response rate • Diagnostic testing procedures applied equally to each group • Appropriate attention to potential confounding variables
Randomized controlled trials (RCTs) and cohort studies	• Initial assembly of comparable groups: For RCTs: adequate randomization, including concealment and whether potential confounders were distributed equally among groups For cohort studies: consideration of potential confounders with either restriction or measurement for adjustment in the analysis; consideration of inception cohorts • Maintenance of comparable groups (includes attrition, crossovers, adherence, contamination) • Important differential loss to follow-up or overall high loss to follow-up • Measurements: equal, reliable, and valid (includes masking of outcome assessment) • Clear definition of interventions • All important outcomes considered • Analysis: adjustment for potential confounders for cohort studies, or intention-to-treat analysis for RCTs
Diagnostic accuracy studies	• Screening test relevant, available for primary care, adequately described • Study uses a credible reference standard, performed regardless of test results • Reference standard interpreted independently of screening test • Handles indeterminate results in a reasonable manner • Spectrum of patients included in study • Sample size • Administration of reliable screening test

To accompany the standard categorization of research design, the third Task Force added a three-category rating of the internal validity of each study: "good," "fair," and "poor." To distinguish among good, fair, and poor, the Task Force modified criteria developed by others[23–26] to create a set of operational parameters for evaluating the internal validity of five different study designs: systematic reviews, case–control studies, RCTs, cohort studies, and diagnostic accuracy studies (Table 3). These criteria are used not as rigid rules but as guidelines; exceptions are made with adequate justification. In general, a good study meets all criteria for that study design; a fair study does not meet all criteria but is judged to have no fatal flaw that invalidates its results; and a poor study contains a fatal flaw.

Thus, the topic team assigns each study two separate ratings: one for study design and one for internal validity. A well-performed RCT, for example, would receive a rating of I-good, whereas a fair cohort study would be rated II-2-fair. In many cases, narrative text is needed to explain the rating of internal validity for the study, especially for those studies that play a pivotal role in the analytic framework. When the quality of an individual study is the subject of significant disagreement, the entire Task Force may be asked to rate the study and the final rating is applied after debate and discussion.

Even well-designed and well-conducted studies may not supply the evidence needed if the studies examine a highly selected population of little relevance to the general population seen in primary care. Thus, external validity—the extent to which the studies reviewed are generalizable to the population of interest—is considered on a par with internal validity. Deciding whether generalizing in specific situation is appropriate is based on explicit principles developed by the Task Force (see Extrapolation and Generalization section).

Evaluating quality at three strata: Stratum 2, the linkage. The quality of evidence in a single study constitutes only one stratum in analyzing the quality of evidence for a preventive service. One might also consider two additional levels of assessment: the quality of the body of evidence for each linkage (key question) in an analytic framework, and the overall quality of the body or bodies of evidence for a preventive service, including all linkages in the analytic framework (Table 4).

In assessing quality at the second level, the body of evidence supporting a given linkage in the analytic framework, the Task Force recognizes three important criteria. The first two follow directly from criteria for the first stratum. Internal validity (including research design) and external validity (generalizability) remain

Table 4. Evaluating the quality of evidence at three strata

Level of evidence	Criteria for judging quality
1. Individual study	• Internal validity[a] • External validity[b]
2. Linkage in the analytic framework	• Aggregate internal validity[a] • Aggregate external validity[b] • Coherence/consistency
3. Entire preventive service	• Quality of the evidence from Stratum 2 for each linkage in the analytic framework • Degree to which there is a complete chain of linkages supported by adequate evidence to connect the preventive service to health outcomes • Degree to which the complete chain of linkages "fit" together[c] • Degree to which the evidence connecting the preventive service and health outcomes is "direct"[d]

[a]Internal validity is the degree to which the study(ies) provides valid evidence for the population and setting in which it was conducted.
[b]External validity is the extent to which the evidence is relevant and generalizable to the population and conditions of typical primary care practice.
[c]"Fit" refers to the degree to which the linkages refer to the same population and conditions. For example, if studies of a screening linkage identify people who are different from those involved in studies of the treatment linkage, the linkages are not supported by evidence that "fits" together.
[d]"Directness" of evidence is inversely proportional to the number of bodies of evidence required to make the connection between the preventive service and health outcomes. Evidence is direct when a single body of evidence makes the connection, and more indirect if two or more bodies of evidence are required.

important, but at this level they are considered in the aggregate for all relevant studies (Table 4).

The third criterion for evaluating the quality of the body of evidence concerning the linkage in an analytic framework is consistency and coherence. Coherence means that a body of evidence makes sense, that is, that the evidence fits together in an understandable model of the situation. The Task Force does not necessarily require consistency, recognizing that studies may produce different results in different populations, and heterogeneity of this sort may still be coherent with the hypothesized model of how interventions relate to outcomes. Consistent results of several studies across different populations and study designs do, however, contribute to coherence.

A topic team considers these three criteria—aggregate internal validity, aggregate external validity, and coherence/consistency—in evaluating the quality of the body of evidence concerning the linkage in an analytic framework (Table 4). It assigns good, fair, or poor ratings to each of these three factors. In making these judgments, the Task Force has no simple formula but rather considers all the evidence, giving greater weight to studies of higher quality. Topic teams write brief explanatory narratives to provide the rationale for their ratings.

Evaluating quality at three strata: Stratum 3, the entire preventive service. The third level of assessing quality considers the evidence for the entire preventive service. Previous Task Forces used the hierarchical rating of research design (Table 2) to describe the best evidence for a preventive service. The evidence for a preventive service would receive a II-2 code, for example, if the best evidence consisted of a controlled cohort study. As noted above, the current USPSTF has added to this grading of research design an assessment of how well the study was conducted.

Even with this addition, however, examination of the analytic framework shows the difficulty in using this rating scheme alone to judge the quality of the evidence for an entire preventive service. The quality of the evidence may depend on which linkage it is examining. For example, the evidence for smoking cessation counseling could be described as grade I-good evidence (because well-performed RCTs have shown that counseling and nicotine replacement therapy reduce smoking rates) or as grade II-2-good evidence (because only cohort studies have shown that stopping smoking improves health). The more precise conceptualization is that smoking cessation counseling consists of multiple components, as reflected in the linkages for its analytic framework (e.g., Figure 2) and that different levels of evidence support each linkage.

The third Task Force adopted an approach that systematically examines the evidence for each linkage, and all linkages together, in the analytic framework. The underlying issue is whether the evidence is adequate to determine the existence and magnitude of a causal connection between the preventive service (on the left side of the analytic framework) and health outcomes (on the right side of the analytic framework).

Rather than applying formal rules for determining the overall quality of evidence, the Task Force adopted a set of general criteria that it considers when making this judgment (Table 4). These criteria are as follows:

- Quality of the evidence from Stratum 2 for each linkage in the analytic framework;
- Degree to which a complete chain of linkages supported by adequate evidence connects the preventive service to health outcomes;
- Degree to which the linkages fit together; and
- Degree to which the evidence connecting the preventive service and health outcomes is direct.

As noted earlier, the directness of evidence is inversely proportional to the number of linkages (bodies of evidence) that must be pieced together to infer that a preventive service has an impact on health. The evidence is most direct if a single body of evidence, corresponding to the overarching linkage in the ana-

lytic framework, provides adequate evidence concerning the existence and magnitude of health effects resulting from the use of the preventive service. The evidence is indirect if, instead of having overarching evidence, one must rely on two or more bodies of evidence corresponding to linkages in the analytic framework to make an adequate connection between the use of the preventive service and health.

Based on these considerations, the Task Force grades the overall quality of the evidence using the same tripartite scheme (good, fair, and poor) applied to other levels of evidence. The Task Force decided against a formal system for assigning these grades. Instead, it makes its reasoning explicit in an explanatory narrative in the recommendation statement, providing the overall assessment of the quality of the evidence and the rationale behind this assessment.

In general, good overall evidence includes a high-quality direct linkage between the preventive service and health outcomes. Fair evidence is typically indirect but it is adequate to complete a chain of linkages across the analytic framework from the preventive service to health outcomes. The evidence is inadequate to make this connection unless the linkages fit together in a meaningful way. For example, in some situations screening may detect people who are different from those involved in studies of treatment efficacy. In this case, the screening and treatment linkages do not fit together. Poor evidence has a formidable break in the evidence chain such that information is inadequate to connect the preventive service and health outcomes.

To make its reasoning explicit, the Task Force includes an explanatory narrative about its overall rating of the evidence in the recommendation statement.

Separating magnitude of effect from quality. When reviewers consider the quality of evidence, they often confound quality of evidence with magnitude of effect. Evidence for an intervention is sometimes described as good if it shows a dramatic effect on outcomes. Strictly speaking, whether a study provides accurate information should be independent of its findings. The magnitude of observed benefits and/or harms from a service, although of critical importance to decisions about whether it should be recommended, is a separate issue from the quality of the data. The Task Force examines magnitude (or effect size) separately from the quality of evidence, but it merges both issues in making its recommendations (see discussion in "Assessing Magnitude of Net Benefit" section).

Assessing Magnitude of Net Benefit

When the overall quality of the evidence is judged to be good or fair, the Task Force proceeds to consider the magnitude of net benefit to be expected from implementation of the preventive service. Determining net benefit requires assessing both the magnitude of benefits and the magnitude of harms and weighing the two. When the evidence is considered to be poor, the Task Force has no scientific basis for making conjectures about magnitude.

The Task Force classifies benefits, harms, and net benefits on a 4-point scale: "substantial," "moderate," "small," and "zero/negative." It has adopted no standardized metric (such as number needed to screen, number needed to treat, number of lives extended, years of life saved, and/or quality-adjusted life years) for comparing net benefit across preventive services. Ideally, a quantitative definition for such terms as substantial or moderate benefit would make these categorizations more defensible, less arbitrary, and more useful to policymakers in ranking the relative priority of preventive services. Unfortunately, the Task Force has not yet solved the methodologic challenges to deriving such a metric.

Although the Task Force has decided against a rigid formula for defining these terms, it has developed a conceptual framework and a process for making these distinctions. In assessing the magnitude of benefits and harms, the Task Force uses a modification of the statistical concept of the confidence interval. The magnitude of effect in individual studies is given by a point estimate surrounded by a confidence interval. Point estimates and confidence intervals often vary among studies of the same question, sometimes considerably. The Task Force examines all relevant studies to construct a general, conceptual "confidence interval" of the range of effect-size values consistent with the literature. It considers the upper and lower bounds of this confidence interval in assessing the magnitude of benefits and harms.

Assessing Magnitude of Benefits

The Task Force thinks of benefit from both population and individual perspectives. For the benefit to be considered substantial, the service must have

- at least a small relative impact on a frequent condition with a substantial population burden, or
- a large impact on an infrequent condition that poses a significant burden at the individual patient level.

For example, counseling for tobacco cessation produces a change in behavior in only a small proportion of patients,[27] but the societal implications are sizable because of the large number of tobacco users in the population and the burden of illness and death that is averted if even a small percentage of people stop smoking. Conversely, phenylketonuria is a grave condition that affects a very small proportion of the population, but neonatal screening markedly reduces morbidity and mortality from the disease.[6] Although the target conditions in these examples differ considerably in

prevalence, the Task Force views both preventive services as having a substantial magnitude of benefit. "Outcomes tables" (similar to "balance sheets"[28]) are the Task Force's standard resource for estimating the magnitude of benefit.[28,29] These tables, prepared by the topic teams for use at Task Force meetings, compare the condition-specific outcomes expected for a hypothetical primary care population with and without use of the preventive service. These comparisons may be extended to consider only people of specified age or risk groups or other aspects of implementation. Thus, outcomes tables allow the Task Force to examine directly how the preventive service affects benefits for various groups.

One important problem with outcomes tables is that the evidence typically differs across table cells. For some services and some groups, the frequency of the outcome may be clear, but for others one can calculate the frequency of the outcome only by making broad assumptions, some with greater scientific support than others. Thus, outcomes tables must provide information about both the frequency of outcomes and how certain we are about that information.

Assessing Magnitude of Harms

The Task Force considers all types of potential harms of a service, both direct harms of the service itself (e.g., those from a screening test or preventive medication) and indirect harms that may be downstream consequences of the initial intervention (e.g., invasive follow-up tests or harms of treatments). The Task Force considers potential medical, psychological, and nonhealth harms (e.g., effects on insurability).

All analytic frameworks include linkages concerning the potential harms of preventive services, and all topic teams search for evidence about these harms. The Task Force strives to give equal weight to benefits and harms in its assessment of net benefit, but the amount of evidence about benefits is usually greater. Few studies provide useful information on adverse outcomes. Thus, the Task Force often finds itself trying to estimate harms based on little evidence. Methods of making this estimation are lacking, but the Task Force continues to discuss ways to frame the range of reasonable estimates of harm for each preventive service.

When evidence on harms is available, the topic teams assess its quality in a manner like that for benefits and include adverse events in the outcomes tables. When few harms data are available, the Task Force does not assume that harms are small or nonexistent. It recognizes a responsibility to consider which harms are likely and to judge their potential frequency and the severity that might ensue from implementing the service.[30] It uses whatever evidence exists to construct a general confidence interval on the 4-point scale (e.g., substantial, moderate, small, and zero/negative) described above.

Assessing Net Benefits: Weighing Benefits and Harms

Value judgments are involved in using the information in an outcomes table to rate either benefits or harms on the Task Force's 4-point scale. Value judgments are also needed to weigh benefits against harms to arrive at a rating of net benefit.

The need to invoke value judgments is most obvious when the Task Force must weigh benefits and harms of different types against each other in coming to a collective assessment of net benefits. For example, although breast cancer screening for certain age groups may reduce deaths from breast cancer,[31] it also increases the number of women who must experience the anxiety of a work-up for a false-positive mammogram.[32] Determining which of the four categories of net benefit to assign to this service depends greatly on the value one places on each outcome.

In making its determinations of net benefit, the Task Force strives to consider what it believes are the general values of most people. It does this with greater confidence for certain outcomes (e.g., death) about which there is little disagreement about undesirability, but it recognizes that the degree of risk people are willing to accept to avert other outcomes (e.g., cataracts) can vary considerably.[33] When the Task Force perceives that preferences among individuals vary greatly, and that these variations are sufficient to make the average trade-off of benefits and harms a "close call," then it will often assign a C recommendation (see below). This recommendation indicates that the decision is likely to be sensitive to individual patients' preferences.

Extrapolation and Generalization

As noted in the "Review of the Evidence" section, the Task Force regularly faces the issue of generalization in determining the quality of evidence. The Task Force makes recommendations intended for the general primary care situation; for this purpose, high-quality evidence is evidence that is relevant and valid for this setting. When studies examine different situations and settings, the issue of generalization arises.

Likewise, the magnitude of the effect of interest to the Task Force is that resulting from implementation in the primary care setting. Calculations based on extrapolation are usually required to estimate the likely magnitude of effect for the primary care situation.

Some degree of extrapolation and generalization is invariably required to use evidence in the research literature to make guidelines for the primary care situation. For some services, the evidence may provide high-quality information about the efficacy of a preventive service in the hands of experts for a specific subpopulation. For others, evidence about efficacy often comes from studies of symptomatic patients who

are more severely ill than patients who would be discovered by screening. Even when good randomized trials of therapeutic efficacy in asymptomatic patients exist (e.g., therapy of lipid disorders), female, elderly, and younger patients may be underrepresented, and eligibility criteria might exclude patients with characteristics that are typical of a general primary care population. Other commonly encountered issues are whether the efficiency of screening in one practice setting can be replicated in other settings and whether efficacy persists or diminishes beyond the length of time usually covered by available studies.

In the absence of good evidence, to what extent can one use reasoned judgments based on assumptions with varying degrees of scientific support to draw conclusions about the potential benefits and harms of a preventive service? The Task Force developed a policy for determining the conditions under which extrapolation and generalization are reasonable. These conditions include:

- biologic plausibility;
- similarities of the populations studied and primary care patients (in terms of risk factor profile, demographics, ethnicity, gender, clinical presentation, and similar factors);
- similarities of the test or intervention studied to those that would be routinely available or feasible in typical practice; and
- clinical or social environmental circumstances in the studies that could modify the results from those expected in a primary care setting.

Judgments about extrapolation and generalization, because they are often matters of policy and subjective judgment rather than hard science, are made by the Task Force and not the EPCs.

Translating Evidence into Recommendations

General Principles

Making recommendations for clinical practice involves considerations that extend beyond scientific evidence. Direct scientific evidence is of pre-eminent interest, but such issues as cost effectiveness, resource prioritization, logistical factors, ethical and legal concerns, and patient and societal expectations should also be considered.

Historically, the Task Force has taken a conservative, evidence-based approach to this process, making recommendations that reflect primarily the state of the evidence and refraining from making recommendations when they cannot be supported by evidence. This is done with the understanding that clinicians and policymakers must still consider additional factors in making their own decisions.[34] The Task Force sees its purpose as providing users with information about the extent to which recommendations are supported by evidence, allowing them to make more informed decisions about implementation.

Another important issue in making recommendations is the amount and quality of evidence required. As evidence is rarely adequate to provide decision makers with completely valid information about all important outcomes for the population of interest, those creating guidelines must consider how far they are willing to generalize from imperfect evidence. As noted in the Extrapolation and Generalization section, the Task Force believes that such generalizations can be made under defined conditions.

The general principles the Task Force follows in making recommendations are outlined in Table 5. Most of these principles have been discussed in other parts of this paper. They involve both the factors considered by the Task Force in making recommendations (e.g., the most salient types of evidence, feasibility, harms, economic costs, and its target population) and the way in which it considers these factors (e.g., the place of subjectivity, the importance of the population perspective, and the extent to which the evidence connects the service with positive net benefits for patients).

Codes and Wording of Statements

As in the past, the Task Force assigns letter codes to its recommendations and uses standardized phrasing for each category of recommendations (Table 6), but the details have changed from previous versions. The original five-letter scheme, which included an E recommendation category that was rarely used,[6] has been replaced with a four-letter scheme that allows only one classification for recommendations *against* routinely providing a preventive service (D).

Previous definitions for letter codes focused on whether the evidence supported "including the preventive service in the periodic health examination." Current thinking is that preventive services should also be delivered in other contexts, such as illness visits. The new wording thus focuses on whether the service should be "routinely provided."

In the past, the Task Force assigned a C code to recommendations with "insufficient evidence to make a recommendation." Previous Task Forces used this code for a wide assortment of circumstances and thus assigned it to a large proportion of the preventive services they reviewed. Evidence could be insufficient because no studies existed, available studies were of poor quality, studies were of reasonable quality but conflicting, or results were consistent but the magnitude of net benefit was small.

The C recommendation, because of its location in the hierarchical ranking of recommendation grades, implies that the service is less worthy of implementation

Table 5. Principles for making recommendations

- Task Force recommendations are evidence based: They require scientific evidence that persons who receive the preventive service experience better health outcomes than those who do not and that the benefits are large enough to outweigh the harms.

  The Task Force emphasizes evidence that directly links the preventive service with health outcomes. Indirect evidence may be sufficient if it supports the principal links in the analytic framework.

  Although the Task Force acknowledges that subjective judgments do enter into the evaluation of evidence and the weighing of benefits and harms, its recommendations are not based largely on opinion.

  The Task Force is explicit about the scientific rationale for its recommendations.

- The outcomes that matter most in weighing the evidence and making recommendations are health benefits and harms.

  In considering potential benefits, the Task Force focuses on absolute reductions in the risk of outcomes that people can feel or care about.

  In considering potential harms, the Task Force examines harms of all types, including physical, psychological, and nonmedical harms that may occur sooner or later as a result of the preventive service.

  Where possible, the Task Force considers the feasibility of future widespread implementation of the preventive service in making recommendations.

  The Task Force generally takes a population perspective in weighing the magnitude of benefits against the magnitude of harms. In some situations, it may recommend a service with a large potential benefit for a small proportion of the population.

  In assessing net benefits, the Task Force subjectively estimates the population's value for each benefit and harm. When the Task Force judges that the perceived balance of benefits and harms is likely to vary substantially within the population, it may abandon general recommendations and suggest shared decision making at the individual level.

- Where possible, the Task Force considers the total economic costs that result from providing a preventive service, both to individuals and to society, in making recommendations, but costs are not the first priority.

  When the Task Force recommends against a preventive service for economic reasons, it states so explicitly.

- The Task Force does not modify its recommendations to accommodate concerns about insurance coverage of preventive services, medicolegal liability, or legislation, but users of the recommendations may need to do so.

- Recommendations apply only to asymptomatic persons or those with unrecognized signs or symptoms of the target condition for which the preventive service is intended. They also apply only to preventive services initiated in the clinical setting.

than services that receive an A or a B recommendation. The current Task Force believes that such pejorative conclusions should be applied only when the evidence provides a basis for inferring that the magnitude of net benefit is smaller than for interventions that merit higher ratings. In other instances, in which evidence is of poor quality or conflicting, the possibility of substantial benefit (or substantial harm) cannot be excluded on scientific grounds and thus the Task Force can make no evidence-based judgments about the service.

To address these cases, the Task Force has created a new recommendation category, the I recommendation (insufficient evidence). It has also intentionally chosen a letter distant from the A–D hierarchy to signal its reluctance to pass judgment about the effectiveness of the interventions that receive this rating. The Task Force gives an I recommendation when studies are lacking or of poor quality or when they produce conflicting results that do not permit conclusions about likely benefits and harms.

For the A–D recommendations, the Task Force has adopted a more formalized process for translating the evidence into group judgments about how strongly to recommend the intervention than had been applied in the past. In earlier years, the simplistic notion was that services supported by RCTs always received A recommendations. The new approach recognizes that the importance of providing the preventive service depends not only on the quality of the evidence but also on the magnitude of net benefit to patients or populations. In an effort to ensure that both dimensions—quality and magnitude—are addressed systematically in assigning letter codes, the Task Force now uses a recommendation grid (Table 7) that makes the process more explicit.

As shown, code A indicates that the quality of evidence is good and the magnitude of net benefits is substantial: The Task Force "strongly recommends" that these services be routinely provided (Table 6). The B code indicates that the Task Force has found that either the quality of the evidence or the magnitude of net benefits (or both) is less than would be needed to warrant an A. Primary care providers should not necessarily give higher priority to A over B services. Setting

Table 6. Standard recommendation language

Recommendation	Language[a]
A	The USPSTF strongly recommends that clinicians routinely provide [the service] to eligible patients. (The USPSTF found good evidence that [the service] improves important health outcomes and concludes that benefits substantially outweigh harms.)
B	The USPSTF recommends that clinicians routinely provide [the service] to eligible patients. (The USPSTF found at least fair evidence that [the service] improves important health outcomes and concludes that benefits outweigh harms.)
C	The USPSTF makes no recommendation for or against routine provision of [the service]. (The USPSTF found at least fair evidence that [the service] can improve health outcomes but concludes that the balance of the benefits and harms is too close to justify a general recommendation.)
D	The USPSTF recommends against routinely providing [the service] to asymptomatic patients. (The USPSTF found at least fair evidence that [the service] is ineffective or that harms outweigh benefits.)
I	The USPSTF concludes that the evidence is insufficient to recommend for or against routinely providing [the service]. (Evidence that [the service] is effective is lacking, of poor quality, or conflicting and the balance of benefits and harms cannot be determined.)

[a]All statements specify the population for which the recommendation is intended and are followed by a rationale statement providing information about the overall grade of evidence and the net benefit from implementing the service.
USPSTF, U.S. Preventive Services Task Force.

priorities for offering, providing, or reimbursing these services should include consideration of time and resource requirements, which are beyond the scope of the Task Force's review. Other groups have undertaken this important work.[35]

The C code indicates that the quality of evidence is either good or fair but that the magnitude of net benefits, as judged in the subjective process outlined above, is too small to make a general recommendation. In these cases, the Task Force "makes no recommendation for or against routinely providing the service." Clinicians and policymakers may choose to offer the service for other reasons—such as considerations other than scientific evidence or because benefits for individual patients are expected to exceed those observed in studies—but the Task Force rating is meant to advise them that existing evidence does not document substantial net benefit for the average patient.

The D code indicates that the evidence is good or fair but that net benefit is probably either zero or negative. In these situations, the Task Force recommends against routine use of the service.

When the evidence is poor, the Task Force cannot distinguish between substantial or moderate net benefits on the one hand and small or zero/negative net benefits on the other. In these cases, the Task Force uses code I to indicate that it cannot make a recommendation for or against routinely providing the service. Because extant evidence cannot yet clarify whether the net benefits of the service are large or small (or negative), this rating advises clinicians and policymakers that determination of whether to provide these services routinely cannot be based on evidence; such decisions must be based on factors other than science.

Table 7. Recommendation grid

Quality of evidence	Net benefit			
	Substantial	Moderate	Small	Zero/negative
Good	A	B	C	D
Fair	B	B	C	D
Poor = I				

Drafting the Report

In its earliest days, background papers and recommendations of the Task Force were written by individual panel members assigned to those topics. In later years, they were written by staff with close oversight by the Task Force. In time a sharp demarcation has evolved between descriptions of the evidence and recommendations.

Thus, for the third Task Force, topic teams led by EPC staff write systematic evidence reviews. These reviews define the strengths and limits of the evidence but stop short of making recommendations.

Systematic evidence reviews typically include the full version (available from AHRQ and accessible on its website, www.ahrq.gov) and a shorter summary such as those published in this issue. As a work product prepared under contract for AHRQ, the systematic evidence reviews must be approved by the agency before public release. The reviews remain pure descriptions of the science; because they are published separately, groups other than the Task Force can use them to formulate their own guidelines and recommendations.

The summary reviews are typically coupled with a "recommendation and rationale" document, written by the Task Force, which contains recommendations and their supporting rationales. Recommendations, which cross the line from science into policy, are based on formal voting procedures that include explicit rules for determining the views of the majority.

The Task Force has an explicit policy concerning conflict of interest. All members and EPC staff disclose

at each meeting if they have an important financial, organizational, or intellectual conflict for each topic being discussed. Task Force members and EPC staff with conflicts can participate in discussions about evidence, but members abstain from voting on recommendations about the topic in question.

Recommendations are independent of the government. They neither require clearance from nor represent the policy of AHRQ or the U.S. Public Health Service, although efforts are made to consult with relevant agencies to reduce unnecessary discrepancies among guidelines.

The Task Force chair or liaisons on the topic team generally compose the first draft of the recommendation and rationale statement, which the full panel then reviews and edits. These statements have the general structure of the chapters in previous editions of the *Guide to Clinical Preventive Services*.[6] Specifically, they include a recommendation statement and code, a rationale statement, and a brief discussion of clinical interventions. The clinical intervention section is meant to provide more specific information and guidance to clinicians about the service, sometimes discussing factors beyond the quality of the evidence and the magnitude of net benefit that must be considered with implementation.

External Review

Before the Task Force makes its final determinations about recommendations on a given preventive service, the EPC and AHRQ send a draft systematic evidence review to four to six external experts and to federal agencies and professional and disease-based health organizations with interests in the topic. They ask the experts to examine the review critically for accuracy and completeness and to respond to a series of specific questions about the document. After assembling these external review comments and documenting the proposed response to key comments, the topic team presents this information to the Task Force in memo form. In this way, the Task Force can consider these external comments and a final version of the systematic review before it votes on its final recommendations about the service.

Conclusion

Methods for making evidence-based practice policies are evolving. At one extreme, guidelines panels could insist on direct evidence or point to any information gaps to justify a negative recommendation for almost any service. Such an approach would result in positive recommendations only for services that had a very narrow confidence interval for net benefit, but many effective services would not be recommended. At the other extreme, guideline groups that accept incomplete data and allow easy extrapolation make many positive recommendations, but they have less certainty that the services they recommend actually produce more benefit than harm.

In avoiding these extremes, the Task Force has wrestled with several gaps in existing methodology for assessing the quality of evidence, for integrating bodies of evidence, and for translating evidence into guidelines. It continues to address several knotty questions: Can criteria for the internal validity of studies be consistently applied across preventive services? How reliable are such criteria in identifying studies with misleading results? How much weight should be given to various degrees of information gaps, particularly those concerning potential harms and generalizations from research studies to everyday practice? Should the Task Force modify any of these methods when dealing with counseling services?

More methodologic research is warranted in several key areas. Principal among these are efforts to determine the best factors to consider in using evidence-based principles to guide judgments about the magnitude of benefits and harms when the available evidence is fair in quality and when gaps exist in the framework supporting effectiveness. These and other challenges will make the methods of the Task Force, like those of other evidence-based guideline programs, a work in progress for many years.

This paper was developed by the Research Triangle Institute–University of North Carolina at Chapel Hill (RTI-UNC) and the Oregon Health Sciences University (OHSU) Evidence-Based Practice Centers under contracts from the Agency for Healthcare Research and Quality (contract nos. 290-97-0011 and 290-97-0018, respectively). We acknowledge the assistance of Jacqueline Besteman, JD, MA, EPC Program Officer; the AHRQ staff working with the third Task Force; and the staffs of the EPCs at RTI-UNC and at OHSU for their many hours of work in support of this effort. We also acknowledge the assistance of the Counseling and Behavioral Issues Work Group of the Task Force, Evelyn Whitlock, MD, MPH, convenor. Finally, we also acknowledge the major contribution of the entire third U.S. Preventive Services Task Force for its support and intellectual stimulation.

The authors of this article are responsible for its contents, including any clinical or treatment recommendations. No statement in this article should be construed as an official position of the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

References

1. Field MJ, Lohr KN, eds. Guidelines for clinical practice: from development to use. Washington, DC: National Academy Press, 1992 (for Institute of Medicine).
2. Woolf SH, George JN. Evidence-based medicine: interpreting studies and setting policy. Hematol Oncol Clin N Amer 2000;14:761–84.
3. Mulrow CD, Cook D, eds. Systematic reviews: synthesis of best evidence for health care decisions. Philadelphia: American College of Physicians, 1998.

4. Cook D, Giacomini M. The trials and tribulations of clinical practice guidelines. JAMA 1999;281:1950–1.
5. Lawrence RS, Mickalide AD, Kamerow DB, Woolf SH. Report of the U.S. Preventive Services Task Force. JAMA 1990;263:436–7.
6. U.S. Preventive Services Task Force. Guide to clinical preventive services: report of the U.S. Preventive Services Task Force, 2nd ed., Washington, DC: Office of Disease Prevention and Health Promotion, U.S. Government Printing Office, 1996.
7. Eddy DM. Clinical decision making: from theory to practice. A collection of essays from JAMA. Boston: Jones and Bartlett Publishers, 1995.
8. Pignone MP, Phillips CJ, Atkins D, Teutsch SM, Mulrow CD, Lohr KN. Screening and treating adults for lipids disorders. Am J Prev Med 2001; 20(suppl 3):77–89.
9. Briss PA, Zaza S, Pappaioanou M, et al. Developing an evidence-based guide to community preventive services: methods. Am J Prev Med 2000; 18(suppl 1):35–43.
10. Meade MO, Richardson WS. Selecting and appraising studies for a systematic review. In: Mulrow CD, Cook D, eds. Systematic reviews: synthesis of best evidence for health care decisions. Philadelphia: American College of Physicians, 1998:81–90.
11. Woolf SH, DiGuiseppi CG, Atkins D, Kamerow DB. Developing evidence-based clinical practice guidelines: lessons learned by the U.S. Preventive Services Task Force. Ann Rev Public Health 1996;17:511–38.
12. Battista RN, Fletcher SW. Making recommendations on preventive practices: methodological issues. Am J Prev Med 1988;4(suppl 4):53–67.
13. Mulrow C, Langhorne P, Grimshaw J. Integrating heterogeneous pieces of evidence in systematic reviews. In: Mulrow CD, Cook D, eds. Systematic reviews: synthesis of best evidence for health care decisions. Philadelphia: American College of Physicians, 1998:103–12.
14. Nelson HD, Helfand M. Screening for chlamydial infection. Am J Prev Med 2001;20(suppl 3):95–107.
15. Helfand M, Mahon SM, Eden KB, Frame PS, Orleans CT. Screening for skin cancer. Am J Prev Med 2001;20(suppl 3):47–58.
16. Wilson JMG, Junger G. Principles and practice of screening for disease. Geneva: World Health Organization, 1968 (Public Health Papers No. 34).
17. Frame PS, Carlson SJ. A critical review of periodic health screening using specific screening criteria. J Fam Pract 1975;2:29–36, 123–9, 189–94, 283–9.
18. Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users' guides to the medical literature. XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. JAMA 1999;282:771–8.
19. Gøtzsche PC, Liberati A, Torri V, Rossetti L. Beware of surrogate outcome measures. Int J Tech Assess Health Care 1996;12:238–46.
20. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. J Qual Improv 1999;25:470–9.
21. Hornberger J, Wrone E. When to base clinical policies on observational versus randomized trial data. Ann Intern Med 1997;127:697–703.
22. Feinstein AR, Horwitz RI. Problems in the "evidence" of "evidence-based medicine." Am J Med 1997;103:529–35.
23. Oxman AD, Cook DJ, Guyatt GH, Evidence-Based Medicine Working Group. Users' guides to the medical literature: how to use an overview. JAMA 1994;272:1367–71.
24. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. J Gen Intern Med 1989; 4:288–295.
25. Guyatt GH, Sackett DL, Cook DJ, Evidence-Based Medicine Working Group. Users' guides to the medical literature. I. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA 1993;270:2598–601.
26. Laupacis A, Wells G, Richardson WS, Tugwell P, Evidence-Based Medicine Working Group. Users' guides to the medical literature V. How to use an article about prognosis. JAMA 1994;272:234–7.
27. Russell MA, Wilson C, Taylor C, Baker CD. Effect of general practitioners' advice against smoking. BMJ 1979;2:231–5.
28. Eddy DM. Comparing benefits and harms: the balance sheet. JAMA 1990;263:2493, 2498, 2501.
29. Braddick M, Stuart M, Hrachovec J. The use of balance sheets in developing clinical guidelines. J Am Board Fam Pract 1999;12:48–54.
30. Ewart RM. Primum non nocere and the quality of evidence: rethinking the ethics of screening. J Am Board Fam Pract 2000;13:188–96.
31. Fletcher SW, Black W, Harris R, Rimer B, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:644–56.
32. Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med 1998;338:1089–96.
33. Nease RF Jr, Kneeland T, O'Connor GT, et al. Variation in patient utilities for outcomes of the management of chronic stable angina: implications for clinical practice guidelines. JAMA 1995;273:1185–90.
34. Woolf SH, Dickey LL. Differing perspectives on preventive care guidelines: a new look at the mammography controversy. Am J Prev Med 1999;17: 260–8.
35. Coffield AB, Maciosek MV, McGinnis JM, et al. Priorities among recommended clinical preventive services. Am J Prev Med 2001. In press.