Update on the Methods of the U.S. Preventive Services Task Force: Estimating Certainty and Magnitude of Net Benefit

George F. Sawaya, MD; Janelle Guirguis-Blake, MD, MPH; Michael LeFevre, MD, MSPH; Russell Harris, MD, MPH; and Diana Petitti, MD, MPH, for the U.S. Preventive Services Task Force

The major goal of the U.S. Preventive Services Task Force (USPSTF) is to provide a reliable and accurate source of evidence-based recommendations on a wide range of preventive services. In this article, the USPSTF updates and reviews the process by which it evaluates evidence, determines the certainty and magnitude of net benefit, and gives a final letter grade to recommendations. Because direct evidence about prevention is often unavailable, the Task Force usually considers indirect evidence. To guide its selection of indirect evidence, a "chain of evidence" is constructed within an analytic framework. The Task Force examines evidence of various research designs that addresses the key questions within the framework. New terms have been added to describe the USPSTF's judgment about the evidence for each key question: "convincing," "adequate," or "inadequate." For increased clarity, the USPSTF has changed its description of overall evidence of net benefit for the preventive service from "good," "fair," or "poor" quality to "high," "moderate," or "low" certainty. This rating considers the extent to which an uninterrupted chain of evidence exists across the analytic framework. Individual studies will continue to be judged as being of "good," "fair," or "poor" quality. Using outcomes tables, the USPSTF estimates the magnitude of benefits and the magnitude of harms, and synthesizes them into an estimate of the magnitude of net benefit. Although some judgment is required at all steps, the USPSTF strives to make the process as explicit and transparent as possible. The USPSTF anticipates that its methods for making evidence-based recommendations will continue to evolve.

Ann Intern Med. 2007;147:871-875. www.annals.org

For author affiliations, see end of text.

The major goal of the U.S. Preventive Services Task Force (USPSTF) is to provide clinicians and policymakers with a reliable and accurate source of evidence-based recommendations on a wide range of preventive services. To accomplish this goal, the USPSTF systematically reviews the evidence concerning both the benefits and harms of widespread implementation of a preventive service. It then assesses the certainty of the evidence and the magnitude of the benefits and harms. On the basis of this assessment, the USPSTF assigns a letter grade to each preventive service signifying its recommendation about provision of the service (**Table 1**).

An important, but often challenging, step is determining the balance between benefits and harms to estimate "net benefit" (that is, benefits minus harms). In this issue, the Task Force reports an update to its recommendation for carotid artery stenosis screening (1, 2). Because carotid artery stenosis screening has both known benefits and harms, estimating net benefit was critical in the final USPSTF recommendation that clinicians not provide carotid artery stenosis screening in asymptomatic people (recommendation letter grade D). Release of this recommendation provides an opportunity for the Task Force to update and explain to a clinical audience the process by which it evaluates evidence, determines the certainty and magnitude of net benefit, and gives a letter grade to the recommendation. We will do this by considering 3 questions: 1) What evidence does the Task Force consider to estimate net benefit? 2) How does the Task Force estimate the certainty of net benefit? and 3) How does the Task Force estimate the magnitude of net benefit?

What Evidence Does the Task Force Consider to Estimate Net Benefit?

The overarching question that the Task Force seeks to answer for every preventive service is whether evidence suggests that provision of the service would improve health outcomes if implemented in a general primary care population. For screening topics, this standard could be met by a large randomized, controlled trial (RCT) in a representative asymptomatic population with follow-up of all members of both the group "invited for screening" and the group "not invited for screening." For example, the Multicentre Aneurysm Screening Study (3) was a population-based RCT of screening for abdominal aortic aneurysm in which 67 800 asymptomatic men age 65 to 74 years in the United Kingdom were randomly assigned to be invited or not to be invited for screening. Both groups were followed for a mean of 4.1 years, and abdominal aortic aneurysm–related mortality and all-cause mortality were compared.

No RCTs of carotid artery stenosis screening have been published; however, RCTs comparing carotid endarterectomy to medical management of asymptomatic carotid artery stenosis are available. The distinction between RCTs that randomly assign people to undergo screening versus RCTs that randomly assign people known to have a condition to an intervention is important. In contrast to the latter, RCTs of screening take into account the false-

See also:

Web-Only
Conversion of graphics into slides

© 2007 American College of Physicians

Table 1. **U.S. Preventive Services Task Force Recommendation Grid***

Certainty of Net Benefit	Magnitude of Net Benefit			
	Substantial	Moderate	Small	Zero/ Negative
High	A	B	C	D
Moderate	B	B	C	D
Low	Insufficient			

* *A*, *B*, *C*, *D*, and *Insufficient* represent the letter grades of recommendation or statement of insufficient evidence assigned by the U.S. Preventive Services Task Force after assessing certainty and magnitude of net benefit of the service.

positive and false-negative rates of the screening test, the possibility of adverse events from the test, the accuracy and potential for adverse events of any subsequent confirmatory diagnostic tests, and the inevitable failure to follow through on the test or any subsequent steps needed before the therapeutic intervention is delivered. In addition, conditions detected by screening may have different biological characteristics than those detected in other ways. The benefits of treating screened individuals, therefore, cannot be assumed to be the same as those of treating symptomatic individuals. Screening trials directly answer a simple question important to the primary care setting: Does screening for a certain condition improve health outcomes?

Direct RCT evidence about screening is often unavailable, so the Task Force considers indirect evidence. To guide its selection of indirect evidence, the Task Force constructs a "chain of evidence" within an analytic framework. Figure 1 of the evidence update (2) in this issue (page 861) shows the analytic framework for the Task Force assessment of carotid artery stenosis screening. Each arrow in the framework defines a key question, and each key question represents a link in the chain of evidence. Rectangles in the framework represent the intermediate outcomes (rounded corners) or the health outcomes (square corners); ovals represent harms. To form an unbroken chain, evidence must support each link in the chain, thereby connecting the target population (far left side of the framework) to the improved health outcome (far right side of the framework).

For each key question, the body of pertinent literature is critically appraised, focusing on 6 questions (**Table 2**). The USPSTF will now describe its judgment about the evidence for each key question as "convincing," "adequate," or "inadequate." Evidence may be considered convincing when derived from several high-quality studies with consistent, logical results that are generalizable to the U.S. primary care population and setting. Evidence may be deemed adequate when, on the basis of judgment, most but not all of these 6 questions are answered favorably. When evidence is conflicting or the studies are of poor quality individually or in aggregate, the evidence for a key question is considered inadequate. Inadequate evidence may create a critical gap in the evidence chain.

How Does the Task Force Estimate the Certainty of Net Benefit?

The next step in the Task Force process is to use the evidence from the key questions to assess whether there would be net benefit if the service were implemented. In 2001, the USPSTF published an article that documented its systematic processes of evidence evaluation and recommendation development (4). At that time, the Task Force's overall assessment of evidence was described as good, fair, or poor. The Task Force realized that this rating seemed to apply only to how well studies were conducted and did not fully capture all of the issues that go into an overall assessment of the evidence about net benefit. To avoid confusion, the USPSTF has changed its terminology. Whereas individual study quality will continue to be characterized as good, fair, or poor, the term *certainty* will now be used to describe the Task Force's assessment of the overall body of evidence about net benefit of a preventive service and the likelihood that the assessment is correct. Certainty will be determined by considering all 6 questions in **Table 2**; the judgment about certainty will be described as high, moderate, or low.

In making its assessment of certainty about net benefit, the evaluation of the evidence from each key question plays a primary role. It is important to note that the Task Force makes recommendations for real-world medical practice in the United States and must determine to what extent the evidence for each key question—even evidence from screening RCTs or treatment RCTs—can be applied to the general primary care population. Frequently, studies are conducted in highly selected populations under special conditions. The Task Force must consider differences between the general primary care population and the populations studied in RCTs and make judgments about the likelihood of observing the same effect in actual practice. For carotid artery stenosis screening, the Task Force searched for evidence about the true prevalence of high-grade carotid artery

Table 2. **Questions Considered by the U.S. Preventive Services Task Force for Evaluating Evidence Related Both to Key Questions and to the Overall Certainty of the Evidence of Net Benefit for the Preventive Service**

1. Do the studies have the appropriate research design to answer the key question(s)?
2. To what extent are the existing studies of high quality? (i.e., what is the internal validity?)
3. To what extent are the results of the studies generalizable to the general U.S. primary care population and situation? (i.e., what is the external validity?)
4. How many studies have been conducted that address the key question(s)? How large are the studies? (i.e., what is the precision of the evidence?)
5. How consistent are the results of the studies?
6. Are there additional factors that assist us in drawing conclusions (e.g., presence or absence of dose–response effects, fit within a biologic model)?

stenosis in the general population, the generalizability of treatment effectiveness estimates based on RCTs conducted in selected populations, and the complication rate from carotid endarterectomy in asymptomatic individuals if performed in nontrial settings (for example, community hospitals).

It is also important to note that 1 of the key questions in the analytic framework refers to the potential harms of the preventive service. The Task Force considers the evidence about the benefits and harms of preventive services separately and equally. Data about harms are often obtained from observational studies because harms observed in RCTs may not be representative of those found in usual practice and because some harms are not completely measured and reported in RCTs. For example, the surgeons who enrolled patients in RCTs of carotid artery stenosis were selected on the basis of their low postoperative stroke and mortality rates. Widespread screening for carotid artery stenosis would invariably lead to surgical treatment provided in hospitals (or by surgeons) with higher rates of complications. The harms of screening for carotid artery stenosis, including the harms from carotid angiography to confirm the diagnosis of carotid artery stenosis in patients screening positive by carotid ultrasonography, were not captured in some treatment RCTs.

Putting the body of evidence for all key questions together as a chain, the Task Force assesses the certainty of net benefit of a preventive service by asking the 6 major questions in **Table 2**. The Task Force would rate a body of convincing evidence about the benefits of a service that, for example, derives from several RCTs of screening in which the estimate of benefits can be generalized to the general primary care population as "high" certainty (**Table 3**). The Task Force would rate a body of evidence that was not clearly applicable to general practice or has other defects in quality, research design, or consistency of studies as "moderate" certainty. Certainty is "low" when, for example, there are gaps in the evidence linking parts of the analytic framework, when evidence to determine the harms of treatment is unavailable, or when evidence about the benefits of treatment is insufficient. **Table 4** summarizes the current terminology used by the Task Force to describe the critical assessment of evidence at all 3 levels: individual studies, key questions, and overall certainty of net benefit of the preventive service.

The Task Force rated the evidence about the net benefits of screening for carotid artery stenosis as being of moderate certainty. Several factors contributed to this rating: No screening RCTs for carotid artery stenosis have been published; the treatment RCTs included patients with characteristics likely to be different from those identified by screening asymptomatic individuals in primary care settings; the surgeons who participated in the treatment trials had complication rates that could not be generalized in usual care settings; and finally, there would probably be important harms from follow-up of positive screening tests—in some instances, angiography. If the Task Force assesses the overall evidence of net benefit as being of high or moderate certainty, then it proceeds to estimate the magnitude of net benefit.

***Table 3.* U.S. Preventive Services Task Force Levels of Certainty Regarding Net Benefit**

Level of Certainty*	Description
High	The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.
Moderate	The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: the number, size, or quality of individual studies inconsistency of findings across individual studies limited generalizability of findings to routine primary care practice lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion.
Low	The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: the limited number or size of studies important flaws in study design or methods inconsistency of findings across individual studies gaps in the chain of evidence findings that are not generalizable to routine primary care practice a lack of information on important health outcomes. More information may allow an estimation of effects on health outcomes.

* The U.S. Preventive Services Task Force (USPSTF) defines *certainty* as "likelihood that the USPSTF assessment of the net benefit of a preventive service is correct." The net benefit is defined as benefit minus harm of the preventive service as implemented in a general primary care population. The USPSTF assigns a certainty level based on the nature of the overall evidence available to assess the net benefit of a preventive service.

How Does the Task Force Estimate the Magnitude of Net Benefit?

The Task Force gives equal attention to benefits and harms because preventive interventions may result in harms as a direct consequence of the service or for other downstream reasons. For example, some patients with a positive screening test for carotid artery stenosis will undergo carotid angiography and may have a stroke from the procedure—a downstream consequence of screening. If the test is falsely positive and the patient does not have confirmatory angiography, he or she may have an unnecessary carotid endarterectomy, possibly leading to stroke as a complication.

The Task Force attempts to quantify the magnitude of benefits and harms that would result from implementing the preventive service in the general primary care population. One way of doing this is by using such measurements as number needed to treat (the number of people who would need to be treated for some defined period to pre-

Table 4. **U.S. Preventive Services Task Force Terminology to Describe the Critical Assessment of Evidence at 3 Levels: Individual Studies, Key Questions, and Overall Certainty of Net Benefit of the Preventive Service**

Level of Evidence Assessed	Terminology	Criteria Used to Select Terminology
Individual studies	Good, fair, poor (quality)	Critical appraisal; judgment
Key questions in analytic framework*	Convincing, adequate, inadequate (evidence)	6 questions in Table 2; judgment
Overall certainty of net benefit of the preventive service	High, moderate, low (certainty)	6 questions in Table 2; judgment

* This terminology is not reflected in the carotid artery stenosis screening recommendation statement in this issue (1), but it will appear in future recommendation statements.

vent 1 adverse health event) or number needed to screen (the number of people who would need to be screened for some defined period to prevent 1 adverse health event). One can also derive a similar number needed to harm (the number of people who would need to be treated or screened for a defined time to cause 1 adverse health event). The Task Force does not have single numbers needed to treat, screen, or harm that it considers a threshold for drawing a conclusion about the magnitude of net benefit.

Once the Task Force estimates the magnitude of benefits and harms, it faces the further challenge of synthesizing these assessments into an estimate of the magnitude of net benefit. Weighing the balance of benefits and harms can be challenging—benefits are often quantified in terms of lives extended or illness events averted, while harms may be measured in terms of the health consequences of false-positive screening tests or adverse effects of treatment. For example, the benefits of prophylactic aspirin therapy among men include fewer coronary heart events, and the harms include more major gastrointestinal bleeding episodes. Although formal decision analyses and cost-effectiveness analyses have been proposed as an objective method to weigh benefits and harms, the Task Force recognizes that such analyses can be complex and opaque and that they may rely on various assumptions, each of which may have substantial uncertainty.

In 2001 (4), the USPSTF introduced the concept of an "outcomes table" to synthesize information on the magnitude of benefits and harms. The outcomes table estimates the actual number of health outcomes (both benefits and harms) for hypothetical groups that do and do not receive the preventive service. Although an outcomes table relies on assumptions and involves uncertainty, the Task Force believes it provides a transparent, objective method to estimate population benefits and harms were the preventive service implemented.

The outcomes table (see Table 2 in the evidence update [2] in this issue [page 868]) was used by the Task Force to arrive at its recommendation for carotid artery stenosis screening. In the best-case scenario (that is, all positive screening test results are confirmed with magnetic resonance angiography, and patients and surgeons are similar to those in the clinical trials), screening, evaluating, and treating 100 000 average-risk individuals would be expected to prevent about 23 strokes over 5 years; or about 1 stroke would be prevented for every 4348 people screened. In a very high-risk population, screening would be expected to prevent about 217 strokes for every 100 000 people screened over a 5-year period. The best-case scenario number needed to screen to prevent 1 stroke, therefore, is estimated at about 461. To assess net benefit, this number must be weighed against the potentially harmful experience (including the time and effort of the patients and clinicians) of the people who do not benefit (number needed to screen minus 1; either 4347 or 460).

Given the results of the outcomes table, the Task Force then categorizes the magnitude of net benefit as being substantial, moderate, small, or zero/negative. This last category refers to preventive measures that, if implemented in the general primary care population, can be expected to achieve no net benefit or to result in overall harm. Despite the quantity of objective evidence reviewed, the Task Force must use judgment in determining final estimates of net benefit; outcomes tables help make this judgment explicit and transparent.

The Task Force can rarely, if ever, assign an exact magnitude to the benefits or harms of implementing a preventive service. It can, however, put boundaries around the estimate of net benefits. The upper and lower boundary limits on the net estimated benefit make up a "conceptual confidence interval." This range is bound by the best- and worst-case scenario estimates based on available evidence. The interval is not meant to have a statistical interpretation. For the carotid artery stenosis screening recommendation, the Task Force "bound" the benefits for screening a primary care population on the basis of population prevalence, screening accuracy, and treatment benefit. Randomized, controlled trials specified the maximum potential benefit from selected individuals having carotid endarterectomies performed by selected surgeons. The Task Force concluded that the magnitude of benefits in the primary care population could not be greater than the magnitude shown in the RCTs and would probably be smaller in real-world settings.

Linking Magnitude and Certainty of Evidence of Net Benefit to Letter Grades

Once the Task Force defines the certainty and magnitude of net benefit, it determines the letter grade that is

linked to its recommendation about provision of the service (**Table 1**). In general, the Task Force believes that preventive services graded as either A or B should be provided to eligible patients. Those with a C grade should not be offered routinely, and D-grade services should not be provided. Services for which the certainty of the evidence is low because of insufficient evidence about net benefit are designated using an I statement, and no recommendation is made. A later article in this series will explain the Task Force's suggestions for clinical practice when current evidence is insufficient.

Wherever possible, the Task Force attempts to provide estimates of the magnitude of net benefit according to underlying disease risk. For example, the Task Force calculated that 1-time screening of a population of 100 000 men age 65 to 74 years for abdominal aortic aneurysm would prevent about 138 abdominal aortic aneurysm–attributable deaths among the 69 000 men who had ever smoked (about 1 in every 500 men screened) and about 17 deaths from abdominal aortic aneurysm among the 31 000 men who had never smoked (about 1 in every 1800 men screened) (5). On the basis of these estimates, the Task Force gave a B recommendation to ever-smoking men and a C recommendation to never-smoking men age 65 to 74 years.

In its final synthesis of all factors relevant to carotid artery stenosis screening in the community setting, the Task Force judged the certainty of net benefit to be moderate. The inability to identify a high-risk population in whom the net benefit was judged to be greater than zero played an important role in judging the magnitude of net benefit as being not greater than zero. The Task Force, therefore, concluded that the benefits of carotid artery stenosis screening in asymptomatic people do not outweigh the harms, a D recommendation.

FUTURE DIRECTIONS

Methods for making evidence-based recommendations continue to evolve. The USPSTF considers the development of evidence on the benefits and harms of a preventive service as a process. As noted in our previous paper (4), 2 methodological extremes could determine when in this process to move from an I statement (no recommendation due to insufficient evidence) to an A, B, C, or D recommendation. One extreme would be to wait to make a recommendation until incontrovertible evidence suggests that benefits do or do not outweigh harms. To do so runs the risk for clinicians acting or not acting while they are waiting for the Task Force recommendation, leading to patient harm by either overuse of an ineffective service or underuse of an effective service. The other extreme would be to make a recommendation too early in the process of gathering evidence or basing the recommendation on weak evidence, running the risk for making a positive recommendation of an ineffective service or a negative recommendation of an effective service, either of which would lead to patient harm through errors of overuse or underuse.

It is the Task Force's goal to avoid both extremes by continually revisiting and reevaluating its methods in light of new advances. In this way, we believe we can be of most help to our audience of clinicians and policymakers.

From the University of California, San Francisco, San Francisco, California; University of Washington, Tacoma, Washington; University of Missouri–Columbia, Columbia, Missouri; University of North Carolina School of Medicine, Chapel Hill, North Carolina; and University of Southern California, Los Angeles, California.

Disclaimer: Recommendations made by the USPSTF are independent of the U.S. government. They should not be construed as an official position of the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

Acknowledgment: The authors thank Tracy Wolff, MD, MPH; Therese Miller, DrPH; and Mary B. Barton, MD, MPP, of the Agency for Healthcare Research and Quality Center for Primary Care, Prevention, and Clinical Partnerships for expert consultation on the innermost mechanics of the USPSTF; and Marion M. Torchia of the Agency for Healthcare Research and Quality Office of Communications and Knowledge Transfer for editorial assistance.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Reprints are available from the USPSTF Web site (www.preventiveservices.ahrq.gov).

Current author addresses are available at www.annals.org.

References

1. **U.S. Preventive Services Task Force.** Screening for carotid artery stenosis: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2007;147:854-9.

2. **Wolff T, Guirguis-Blake J, Miller T, Gillespie M, Harris R.** Screening for carotid artery stenosis: an update of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 2007;147:860-70.

3. **Ashton HA, Buxton MJ, Day NE, Kim LG, Marteau TM, Scott RA, et al.** The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm screening on mortality in men: a randomised controlled trial. Lancet. 2002;360:1531-9. [PMID: 12443589]

4. **Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al.** Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20:21-35. [PMID: 11306229]

5. **Fleming C, Whitlock EP, Beil TL, Lederle FA.** Screening for abdominal aortic aneurysm: a best-evidence systematic review for the U.S. Preventive Services Task Force. Ann Intern Med. 2005;142:203-11. [PMID: 15684209]