Guide to CLINICAL PREVENTIVE SERVICES

Report of the
U.S. Preventive Services
Task Force

SECOND EDITION

1996

Williams & Wilkins
BALTIMORE • PHILADELPHIA • HONG KONG
LONDON • MUNICH • SYDNEY • TOKYO
A WAVERLY COMPANY

97. Gohagan JK, Prorok PC, Kramer BS, et al. Prostate cancer screening in the Prostate, Lung, Colorectal, Ovarian Cancer Screening Trial of the National Cancer Institute. J Urol 1994;152:1905–1909.
98. Department of Health and Human Services, Public Health Service, Office of Disease Prevention and Health Promotion. Put Prevention into Practice education and action kit. Washington, DC: Government Printing Office, 1994.
99. Woolf SH, Jonas S, Lawrence RS, eds. Health promotion and disease prevention in clinical practice. Baltimore: Williams & Wilkins, 1995.
100. Hayward RSA, Steinberg EP, Ford DE, et al. Preventive care guidelines: 1991. Ann Intern Med 1991; 114:758–783.

ii. Methodology

DCEA: See esp. pp. xlvii–li

This report presents a systematic approach to evaluating the effectiveness of clinical preventive services. The recommendations, and the review of evidence from published clinical research on which they are based, are the product of a methodology established at the outset of the project. The intent of this analytic process has been to provide clinicians[a] with current and scientifically defensible information about the effectiveness of different preventive services and the quality of the evidence on which these conclusions are based. This information is intended to help clinicians who have limited time to select the most appropriate preventive services to offer in a periodic health examination for patients of different ages and risk categories. The critical appraisal of evidence is also intended to identify preventive services of uncertain effectiveness as well as those that could result in more harm than good if performed routinely by clinicians.

For the content of this report to be useful, and to clarify differences between the U.S. Preventive Services Task Force recommendations and those of other groups, it is important to understand the process by which this report was developed, as well as how it differs from the consensus development process used to derive many other clinical practice guidelines. First, the objectives of the review process, including the types of preventive services to be examined and the nature of the recommendations to be developed, were carefully defined early in the process. Second, the Task Force adopted explicit criteria for recommending the performance or exclusion of preventive services and applied these "rules of evidence" systematically to each topic it studied. Third, literature searches and assessments of the quality of individual studies were conducted in accordance with rigorous, predetermined methodologic criteria. Fourth, guidelines were adopted for translating these findings into sound clinical practice recommendations. Fifth, these recommendations were reviewed extensively by content experts in the U.S., Canada, Europe, and Australia. Finally, the review comments were examined by the Task Force and a final vote on recommendations was

[a] The provider of preventive services in primary care is often a physician. The term "clinician" is used in this report, however, to include other primary care providers such as nurses, nurse practitioners, physicians' assistants, and other allied health professionals. Although physicians may be better qualified than other providers to perform certain preventive services or to convince patients to change behavior, some preventive services may be more effectively performed by others with special training (e.g., nurses, dietitians, smoking cessation counselors, mental health professionals).

conducted. The hallmarks of this process are that it is evidence-based and explicit. Each step is examined in greater detail below.

Definition of Objectives

Systematic rules were used to select the target conditions and candidate preventive interventions to be evaluated by the Task Force.

Selection of Target Conditions. In the first edition of this report, the Task Force identified 60 of the leading causes of death and disability in the U.S. that were potentially preventable through clinical interventions. This second edition examines most of the same conditions but also reviews evidence regarding new topics that were added to the list in recent years. The new topics were selected by a rank-order process in which topics were graded on the basis of the frequency and severity of the disease and the potential impact of preventive interventions on health outcomes. In general, the Task Force judged the importance of candidate topics on the basis of two criteria:

Burden of Suffering from the Target Condition. This report examines conditions that are relatively common in the U.S. and are of major clinical significance. Thus, consideration was given to both the *prevalence* (proportion of the population affected) and *incidence* (number of new cases per year) of the condition. Conditions that were once common but have become rare because of effective preventive interventions (e.g., poliomyelitis) were included in the review.

Potential Effectiveness of the Preventive Intervention. Conditions were excluded from analysis if the panel could not identify a potentially effective preventive intervention that might be performed by clinicians.

A number of important prevention topics have not yet been examined by the Task Force due to resource and time constraints. The absence of a discussion of these topics in this report does not imply a judgment about their relative importance or effectiveness.

Selection of Preventive Services. For each target condition, the Task Force used two criteria to select the preventive services to be evaluated. First, in general, only preventive services carried out on *asymptomatic persons*[b] were

[b] The term "asymptomatic person" as used in this report differs from its customary meaning in medical practice. Although "asymptomatic" is often considered synonymous with "healthy," the term is used in this report to describe individuals who lack clinical evidence of the target condition. Signs and symptoms of illnesses unrelated to the target condition may be present without affecting the designation of "asymptomatic." Thus, a 70-year-old man with no genitourinary symptoms who is screened for prostate cancer would be designated asymptomatic for that condition, even if he were hospitalized for (unrelated) congestive heart failure. Preventive services recommended for "asymptomatic patients" therefore need not be delivered only during preventive checkups of healthy persons but apply equally to clinical encounters with patients being seen for other reasons (see Chapter iii).

reviewed. Thus, only primary and secondary preventive measures were addressed. In a clinical setting, *primary preventive measures* are those provided to individuals to prevent the onset of a targeted condition (e.g., routine immunization of healthy children), whereas *secondary preventive measures* identify and treat asymptomatic persons who have already developed risk factors or preclinical disease but in whom the condition has not become clinically apparent. Obtaining a Papanicolaou smear to detect cervical dysplasia before the development of cancer and screening for high blood pressure are forms of secondary prevention. Preventive measures that are part of the treatment and management of persons with clinical illnesses, such as cholesterol reduction in patients with coronary heart disease or insulin therapy to prevent the complications of diabetes mellitus, are usually considered *tertiary prevention* and are outside the scope of this report.

The second criterion for selecting preventive services for review was that the maneuver had to be performed in the *clinical setting.* Only those preventive services that would be carried out by clinicians in the context of routine health care were examined. Findings should not be extrapolated to preventive interventions performed in other settings. Screening tests are evaluated in terms of their effectiveness when performed during the clinical encounter (i.e., *case finding*). Screening tests performed solely at schools, work sites, health fairs, and other community locations are generally outside the scope of this report. Also, preventive interventions implemented outside the clinical setting (e.g., health and safety legislation, mandatory screening, community health promotion) are not specifically evaluated, although clinicians can play an important role in promoting such programs and in encouraging the participation of their patients. References to these types of interventions are made occasionally in sections of this book.

Preventive services were divided into three categories: screening tests, counseling interventions, and immunizations and chemoprophylaxis. *Screening tests* are those preventive services in which a special test or standardized examination procedure is used to identify patients requiring special intervention. Nonstandardized historical questions, such as asking patients whether they smoke, and tests involving symptomatic patients are not considered screening tests in this report. *Counseling* interventions are those in which the patient receives information and advice regarding personal behaviors (e.g., diet) that could reduce the risk of subsequent illness or injury. The Task Force did not consider counseling that addresses the health-related behaviors of persons who have already developed signs and symptoms of the target condition. *Immunizations* discussed in this report include vaccines and immunoglobulins (passive immunization) taken by persons with no evidence of infectious disease. *Chemoprophylaxis* as primary prevention refers to the use of drugs or biologics taken by asymptomatic persons to reduce the risk of developing a disease.

Criteria for Determining Effectiveness

Preventive services must meet predetermined criteria to be considered effective. The criteria of effectiveness for the four categories of preventive services (Table 1) provided the analytic framework for the evaluation of effectiveness in the 70 chapters in this report. Each of these criteria must be satisfied to evaluate the "causal pathway"[1] of a preventive service, the chain of events that must occur for a preventive maneuver to influence health outcomes. Thus, a screening test is not considered effective if it lacks sufficient accuracy to detect the condition earlier than without screening or if there is inadequate evidence that early detection improves health outcomes. Similarly, counseling interventions cannot be considered effective in the absence of firm evidence that changing personal behavior can improve outcome and that clinicians can influence this behavior through counseling. Effective immunization and chemoprophylactic regimens require evidence of biologic efficacy; in the case of chemoprophylactic agents, evidence is also necessary that patients will comply with long-term use of the drug.

The methodologic issues involved in evaluating screening tests require further elaboration. As mentioned above, a screening test must satisfy two major requirements to be considered effective:

- The test must be able to detect the target condition earlier than without screening and with sufficient accuracy to avoid producing large numbers of false-positive and false-negative results (*accuracy of screening test*).
- Screening for and treating persons with early disease should improve the likelihood of favorable health outcomes (e.g., disease-specific morbidity or mortality) compared to treating patients when they present with signs or symptoms of the disease.

These two requirements of screening are essential and therefore appear as headings in each of the 53 screening chapters in this report.

Table 1.
Criteria of Effectiveness

Screening tests
• Accuracy of screening tests
• Effectiveness of early detection
Counseling interventions
• Efficacy of risk reduction
• Effectiveness of counseling
Immunizations
• Efficacy of vaccine
Chemoprophylaxis
• Efficacy of chemoprophylaxis
• Effectiveness of counseling

Accuracy of Screening Tests. The "accuracy of a screening test" is used in this report to describe accuracy and reliability. *Accuracy* is measured in terms of two indices: sensitivity and specificity (Table 2). *Sensitivity* refers to the proportion of persons with a condition who correctly test "positive" when screened. A test with poor sensitivity will miss cases (persons with the condition) and will produce a large proportion of *false-negative* results; true cases will be told incorrectly that they are free of disease. *Specificity* refers to the proportion of persons without the condition who correctly test "negative" when screened. A test with poor specificity will result in healthy persons being told that they have the condition (*false positives*). An accepted reference standard ("gold standard") is essential to the empirical determination of sensitivity and specificity, because it defines whether the disease is present and therefore provides the means for distinguishing between "true" and "false" test results.

The use of screening tests with poor sensitivity and/or specificity is of special significance to the clinician because of the potentially serious consequences of false-negative and false-positive results. Persons who receive false-negative results may experience important delays in diagnosis and treatment. Some might develop a false sense of security, resulting in inadequate attention to risk-reducing behaviors and delays in seeking medical care when warning symptoms become present.

Table 2.
Definition of Terms

Term	Definition	Formula[a]
Sensitivity	Proportion of persons with condition who test positive	$\frac{a}{a+c}$
Specificity	Proportion of persons without condition who test negative	$\frac{d}{b+d}$
Positive predictive value	Proportion of persons with positive test who have condition	$\frac{a}{a+b}$
Negative predictive value	Proportion of persons with negative test who do not have condition	$\frac{d}{c+d}$

[a]Explanation of symbols

	Condition Present	Condition Absent
Positive test	a	b
Negative test	c	d

Legend:
a = true positive
b = false positive
c = false negative
d = true negative

False-positive results can lead to follow-up testing that may be uncomfortable, expensive, and, in some cases, potentially harmful. If follow-up testing does not disclose the error, the patient may even receive unnecessary treatment. There may also be psychological consequences. Persons informed of an abnormal medical test that is falsely positive may experience unnecessary anxiety until the error is corrected. Labeling individuals with the results of screening tests may affect behavior; for example, studies have shown that some persons with hypertension identified through screening may experience altered behavior and decreased work productivity.[2,3]

A proper evaluation of a screening test result must therefore include a determination of the likelihood that the patient has the condition. This is done by calculating the *positive predictive value* (PPV) of test results in the population to be screened (Table 2). The PPV is the proportion of positive test results that are correct (true positives). For any given sensitivity and specificity, the PPV increases and decreases in accordance with the prevalence of the target condition in the screened population. If the target condition is sufficiently rare in the screened population, even tests with excellent sensitivity and specificity can have low PPV in these settings, generating more false-positive than true-positive results. This mathematical relationship is best illustrated by an example (see Table 3):

> A population of 100,000 in which the prevalence of a hypothetical cancer is 1% would have 1,000 persons with cancer and 99,000 without cancer. A screening test with 90% sensitivity and 90% specificity would detect 900 of the 1,000 cases, but would also mislabel 9,900 healthy persons. Thus, the PPV (the proportion of persons with positive test results who actually had cancer) would be 900/10,800, or 8.3%. If the same test were performed in a population with a cancer prevalence of 0.1%, the PPV would fall to 0.9%, a ratio of 111 false positives for every true case of cancer detected.

Reliability (reproducibility), the ability of a test to obtain the same result when repeated, is another important consideration in the evaluation of screening tests measuring continuous variables (e.g., cholesterol level). A test with poor reliability, whether due to differences in results obtained by different individuals or laboratories (*interobserver variation*) or by the same observer (*intraobserver variation*), may produce results that vary widely from the correct value, even though the average of the results approximates the true value.

Effectiveness of Early Detection. Even if the test accurately detects early-stage disease, one must also question whether there is any benefit to the patient in having done so. Early detection should lead to the implementation of clinical interventions that can prevent or delay progression of the disorder. Detection of the disorder is of little clinical value if the condition is not

Table 3.
Positive Predictive Value (PPV) and Prevalence

Testing Conditions
Size of population = 100,000
Sensitivity of test = 90%
Specificity of test = 90%

Cancer Prevalence = 1%

	Cancer Present	Cancer Absent
Positive test	900	9,900
Negative test	100	89,100

PPV = 8.3%

Cancer Prevalence = 0.1%

	Cancer Present	Cancer Absent
Positive test	90	9,990
Negative test	10	89,910

PPV = 0.9%

treatable. Thus, *treatment efficacy* is fundamental for an effective screening test. Even with the availability of an efficacious form of treatment, *early detection* must offer added benefit over conventional diagnosis and treatment if screening is to improve outcome. The effectiveness of a screening test is questionable if asymptomatic persons detected through screening have the same health outcome as those who seek medical attention because of symptoms of the disease. Studies of the effectiveness of cancer screening tests, for example, can be influenced by lead-time and length biases.

Lead-Time and Length Bias. It is often difficult to determine with certainty whether early detection truly improves outcome, an especially common problem when evaluating cancer screening tests. For most forms of cancer, 5-year survival is higher for persons identified with early-stage disease. Such data are often interpreted as evidence that early detection of cancer is effective, because death due to cancer appears to be delayed as a result of screening and early treatment. Survival data do not constitute true proof of benefit, however, because they are easily influenced by *lead-time bias*: survival can appear to be lengthened when screening simply advances earlier the time of diagnosis, lengthening the period of time between diagnosis and death without any true prolongation of life.[4]

Length bias can also result in unduly optimistic estimates of the effectiveness of cancer screening. This term refers to the tendency of screening to detect a disproportionate number of cases of slowly progressive disease and to miss aggressive cases that, by virtue of rapid progression, are present in the population only briefly. The "window" between the time a cancer can be detected by screening and the time it will be found because of symptoms is shorter for rapidly growing cancers, so they are less likely to be found by screening. As a result, persons with aggressive malignancies will be under-

represented in the cases detected by screening, and the patients found by screening may do better than unscreened patients even if the screening itself does not influence outcome. Due to this bias, the calculated survival of persons detected through screening could overestimate the actual effectiveness of screening.[4]

Assessing Population Benefits. Although these considerations provide necessary information about the clinical effectiveness of preventive services, other factors must often be examined to obtain a broader picture of the potential health impact on the population as a whole. Interventions of only minor effectiveness in terms of *relative risk* may have significant impact on the population in terms of *attributable risk* if the target condition is common and associated with significant morbidity and mortality. Under these circumstances, a highly effective intervention (in terms of relative risk) that is applied to a small high-risk group may save fewer lives than one of only modest clinical effectiveness applied to large numbers of affected persons (see Table 4). Failure to consider these epidemiologic characteristics of the target condition can lead to misconceptions about overall effectiveness.

Potential adverse effects of interventions must also be considered in assessing overall health impact, but often these effects receive inadequate attention when effectiveness is evaluated. For example, the widely held belief that early detection of disease is beneficial leads many to advocate screening even in the absence of definitive evidence of benefit. Some may discount the clinical significance of potential adverse effects. A critical examination will often reveal that many kinds of testing, especially among ostensibly healthy persons, have potential direct and indirect adverse effects. Direct physical complications from test procedure (e.g., colonic perforation during sigmoidoscopy), labeling and diagnostic errors based on test results (see above), and increased economic costs are all potential consequences of screening tests. Resources devoted to costly screening programs of uncertain effectiveness may consume time, personnel, or money needed for other more effective health care services. To the USPSTF, potential adverse effects are considered clinically relevant and are always

Table 4.
Effective of Mortality Rate on Total Deaths Prevented

Reduction in Mortality with Intervention	Deaths per Year from Target Condition	Total Deaths Prevented with Intervention
50%	10	5
1%	100,000	1,000

evaluated along with potential benefits in determining whether a preventive service should be recommended.

Methodology for Reviewing Evidence

In evaluating effectiveness, the Task Force used a systematic approach to collect evidence from published clinical research and to judge the quality of individual studies.

Literature Retrieval Methods. Studies were obtained for review by searching MEDLARS, the National Library of Medicine computerized information system, primarily using MEDLINE (a bibliographic database of published biomedical journal articles); other MEDLARS databases such as AIDSLINE and CANCERLIT were occasionally used. Searches for some topics involved the PSYCHINFO database and other relevant sources. Searches were generally restricted to English-language publications. Keywords used in the searches are available for most topics. The reference list was supplemented by citations obtained from experts and from reviews of bibliographic listings, textbooks, and other sources. Literature reviews for this report were generally completed before May 1995, and studies published or entered in MEDLARS subsequently are not addressed.

Exclusion Criteria. Many preventive services involve tests or procedures that are not used exclusively in the context of primary or secondary prevention. Sigmoidoscopy, for example, is also performed for purposes other than screening. Thus, studies evaluating the effectiveness of procedures or tests involving patients who are symptomatic or have a history of the target condition were generally not considered admissible evidence for evaluating effectiveness in asymptomatic persons. Such tests were instead considered *diagnostic tests*, even if they were described by investigators as "screening tests." Uncontrolled studies, comparisons between time and place (ecologic or cross-cultural studies, studies with historical controls), descriptive data, and animal studies were generally excluded from the review process when evidence from randomized controlled trials, cohort studies, or case-control studies (see below) was available. *Etiologic evidence* demonstrating a causal relationship between a risk factor and a disease was considered less persuasive than evidence from well-designed *intervention studies* that measure the effectiveness of modifying the risk factor. As mentioned above, studies of preventive interventions not performed by clinicians were generally excluded from review.

Evaluating the Quality of the Evidence. The methodologic quality of individual studies has received special emphasis in this report. Although all types of evidence were considered, greater weight was given to well-de-

signed studies. Studies that examined health outcomes (e.g., measures of morbidity or mortality) were considered more relevant to assessing effectiveness than studies that used intermediate or physiologic outcome measures to infer effectiveness. (Intermediate outcomes, such as changes in blood cholesterol levels, are often associated with, or precede, health outcomes, but their presence or absence does not necessarily prove an effect on health outcomes.) In addition, study designs were given greater weight if they were less subject to confounding (effects on outcomes due to factors other than the intervention under investigation). Three types of study designs received special emphasis: controlled trials, cohort studies, and case-control studies.

In *randomized controlled trials*, participants are assigned randomly to a study group (which receives the intervention) or a control group (which receives a standard treatment, which may be no intervention or a placebo). In this way, all confounding variables, known and unknown, should be distributed randomly and, in general, equally between the study and control groups. Randomization thereby enhances the comparability of the two groups and provides a more valid basis for inferring that the intervention caused the observed outcomes. In a *blinded* trial, the investigators, the subjects, or both (*double-blind study*) are not told to which group subjects have been assigned, so that this knowledge will not influence their assessment of outcome. Controlled trials that are not randomized are more subject to biases, including *selection bias*: persons who volunteer or are assigned by investigators to study groups may differ systematically in characteristics other than the intervention itself, thereby limiting the internal validity and generalizability of the results.

A *cohort study* differs from a clinical trial in that the investigators do not determine at the outset which persons receive the intervention or exposure. Rather, persons who have already been exposed to the risk factor or intervention and controls who have not been exposed are selected by the investigators to be followed *longitudinally* over time in an effort to observe differences in outcome. The Framingham Heart Study, for example, is a large ongoing cohort study providing longitudinal data on cardiovascular disease in residents of a Massachusetts community in whom potential cardiovascular risk factors were first measured nearly 50 years ago. Cohort studies are therefore *observational*, whereas clinical trials are *experimental*. Cohort studies are more subject to systematic bias than randomized trials because treatments, risk factors, and other covariables may be chosen by patients or physicians on the basis of important (and often unrecognized) factors that may affect outcome. It is therefore especially important for investigators to identify and correct for *confounding variables*, related factors that may be more directly responsible for health outcome than the intervention/exposure in question. For example, increased mortality among

persons with low body weight can be due to the confounding variable of underlying illness. Unlike randomized controlled trials, a shortcoming of cohort studies is that one can correct only for known confounding variables.

Both cohort studies and clinical trials have the disadvantage of often requiring large sample sizes and/or many years of observation to provide adequate *statistical power* to measure differences in outcome. Failure to demonstrate a significant effect in such studies may be the result of statistical properties of the study design rather than a true reflection of poor clinical effectiveness. Both clinical trials and cohort studies have the advantages, however, of generally being *prospective* in design—the health outcome is not known at the beginning of the study and therefore is less likely to influence the collection of data—and of better collection of data to ensure the comparability of intervention and control groups.

Large sample sizes and lengthy follow-up periods are often unnecessary in *case-control studies*. This type of study differs from cohort studies and clinical trials in that the study and control groups are selected on the basis of whether they have the *disease* (cases) rather than whether they have been *exposed* to a risk factor or clinical intervention. The design is therefore *retrospective*, with the health outcome already known at the outset. In contrast to the Framingham Heart Study, a case-control study might first identify persons who have suffered myocardial infarction (cases) and those who have not (controls) and evaluate both groups to assess differences in exposure to an agent (e.g., aspirin) that purportedly reduces the risk of myocardial infarction. In case-control studies of cancer screening, prior exposure to a cancer screening test is compared between patients with cancer (cases) and those without (controls). Principal disadvantages of this study design are that important confounding variables may be difficult to identify and adjust for, health outcome is already known and may influence the measurement and interpretation of data (*observer bias*), participants may have difficulty in accurately recalling past medical history and previous exposures (*recall bias*), and improperly selected control groups may invalidate conclusions about the presence or absence of statistical associations. Both case-control and cohort studies are subject to selection biases because patients who engage in preventive behaviors (or who are selected by clinicians to receive preventive services) may differ in important ways from the general population.

Other types of study designs, such as ecologic or cross-national studies, uncontrolled cohort studies, and case reports, can provide useful data but do not generally provide strong evidence for or against effectiveness. Cross-cultural comparisons can demonstrate differences in disease rates between populations or countries, but these differences could be due to a variety of genetic and environmental factors other than the variable in question.

Uncontrolled studies may demonstrate impressive treatment results or better outcomes than have been observed in the past (historical controls), but the absence of internal controls raises the question of whether the results would have occurred even in the absence of the intervention, perhaps as a result of other concurrent medical advances or changes in case selection. For further background on methodologic issues in evaluating clinical research, the reader is referred to other publications.[4–6]

In summary, claims of effectiveness in published research must be interpreted with careful attention to the type of study design. Impressive findings, even if reported to be statistically significant, may be an artifact of measurement error, the manner in which participants were selected, or other design flaws rather than a reflection of a true effect on health outcome. In particular, the p-value, which expresses the probability that a finding could have occurred by chance, does not account for bias. Thus, even impressively low p-values are of little value when the data may be subject to substantial bias. Conversely, research findings suggesting ineffectiveness may result from low statistical power, inadequate follow-up, and other design limitations. A study with inadequate statistical power may fail to demonstrate a significant effect on outcomes because of inadequate sample size rather than because of the limitations of the intervention.

The quality of the evidence is therefore as important as the results. For these reasons, the Task Force used a hierarchy of evidence in which greater weight was given to those study designs that are, in general, less subject to bias and misinterpretation. The hierarchy ranked the following designs in decreasing order of importance: randomized controlled trials, nonrandomized controlled trials, cohort studies, case-control studies, comparisons between time and places, uncontrolled experiments, descriptive studies, and expert opinion. For each of the preventive services examined in this report, the Task Force assigned "evidence ratings" reflecting this hierarchy using a five-point scale (I, II-1, etc.) adapted from the scheme developed originally by the Canadian Task Force on the Periodic Health Examination (see Appendix A).

Due to resource constraints, the Task Force generally did not perform meta-analysis or decision analysis to examine the data or to synthesize the results of multiple studies. For topics in which these techniques are appropriate, the Task Force encourages other groups to conduct such analyses. Previously published meta-analyses or decision analytic models were reviewed by the Task Force in its examination of the literature but generally did not provide the sole basis for its recommendations unless the quality of the studies and analytic model was high.

Updating the Evidence. Because the first edition of the *Guide* reviewed most of the relevant supporting evidence published before 1989, the Task Force adopted an updating process to identify important evidence and new preventive technologies to address in this edition of the report. Literature review and updating of some topics for which little new evidence had been published since 1989 were conducted off-site at academic medical centers under the supervision of Task Force members. Updating of most other topics was performed by research staff at the Office of Disease Prevention and Health Promotion.

Updating was also coordinated with the Canadian Task Force on the Periodic Health Examination, which used a similar methodology to evaluate the effectiveness of preventive services and produced a report with similar format for its Canadian audience.[7] For a number of topics in which differences in population characteristics were not important, draft chapters developed by the Canadian panel were adapted by the U.S. Task Force for inclusion in this report. The chapters on screening for ovarian cancer and hormone replacement therapy (Chapters 14 and 68, respectively) were based on reviews conducted for the American College of Physicians.

Translating Science into Clinical Practice Recommendations

Recommendations to perform or not perform a preventive service can be influenced by multiple factors, including scientific evidence of effectiveness, burden of suffering, costs, and policy concerns. The recommendations in this report are influenced largely by only one factor, scientific evidence, recognizing that the other factors often need to be considered (see below). Task Force recommendations are graded on a five-point scale (A-E), reflecting the strength of evidence in support of the intervention (see Appendix A). Interventions that have been proved effective in well-designed studies or have demonstrated consistent benefit in a large number of studies of weaker design are generally recommended in this report as "A" or "B" recommendations. Interventions that have been proved to be ineffective or harmful are generally not recommended and are assigned "D" or "E" recommendations. Even when there is no definitive evidence that a preventive service is ineffective, a "D" recommendation may be applied if there is no proven benefit and there is a known risk of complications or adverse effects from the preventive maneuver or from the diagnostic and treatment interventions that it generates. Under these conditions of uncertain benefit and known harm, the Task Force often discourages routine performance in the asymptomatic population but recognizes that future research may later establish a favorable benefit-harm relationship that supports routine performance.

For many preventive services (and much of medical practice), there is insufficient evidence that the maneuver is or is not effective in improving outcomes ("C" recommendation). *This lack of evidence of effectiveness does not*

constitute evidence of ineffectiveness. A preventive service can lack evidence and receive a "C" recommendation because no effectiveness studies have been performed. In other cases, studies may have been performed but they may have produced conflicting results. Studies showing no benefit may lack adequate statistical power, making it unclear whether the maneuver would be proved effective if it were tested with a larger sample size. Studies showing a benefit may suffer from other design flaws (e.g., confounding variables) that raise questions about whether the observed effect was due to the experimental intervention or other factors.

In all of these instances, the Task Force gives the preventive service a "C" recommendation, noting that there is insufficient scientific evidence to conclude whether the maneuver should or should not be performed routinely. Practitioners and policy makers often need to consider factors other than science, however, in deciding how to proceed in the absence of evidence. The first of these considerations is potential harm to the patient. In the absence of proven benefit, many would consider the performance of potentially harmful preventive services (e.g., aspirin prophylaxis in pregnancy) to be inappropriate ("primum non nocere"). It may be entirely appropriate, however, to perform preventive services that are essentially harmless if they have a reasonable likelihood of helping the patient (e.g., patient education and counseling). Similar considerations apply to costs. Performing costly preventive services in the absence of evidence (e.g., home uterine activity monitoring for preterm labor) must be viewed differently from inexpensive maneuvers of unproven benefit (e.g., palpating the testicles in young men).

The burden of suffering from the target condition may justify the performance of preventive services, even in the absence of evidence, and similar considerations may apply to an individual patient's risk status. Unproven preventive services that are inappropriate for the general population may be appropriate to consider for individuals at markedly increased risk of the disease. Patient preferences, which are important in all clinical decisions, are essential to consider when contemplating the performance of preventive services of unproven effectiveness. The clinician's responsibility is to provide the patient with the best available information about the potential benefits and harms of the preventive service and to delineate what is known and not known about the probability of these outcomes. Patients can then make informed decisions about which option is appropriate, based on the relative importance that they assign to these outcomes.

These additional considerations account for the different language used by the Task Force in its wording of "C" recommendations. Although all preventive services in the "C" category are identified as having insufficient evidence to recommend for or against the maneuver, the Task Force often adds that arguments for or against the practice can be made on

"other grounds." These include the absence of significant harm or cost, the potential of improving individual or public health, legal requirements ("other grounds" *for* performing the preventive service) and concerns that the potential harms and costs of the maneuver outweigh its potential benefits ("other grounds" for *not* performing the preventive service). In some cases, the Task Force maintains a completely neutral position, stating only that there is insufficient evidence to make a recommendation. The statement that "recommendations may be made on other grounds" is intended to call attention to factors that may help guide the clinical practice; *it does not constitute an explicit recommendation of the Task Force that these services be provided or omitted routinely in the absence of evidence of effectiveness.* Individual clinical decisions should be made on a case-by-case basis.

In selected situations, even preventive services of proven efficacy may not be recommended due to concerns about feasibility and compliance. Benefits observed under carefully controlled experimental conditions may not be generalizable to normal medical practice. That is, the preventive service may have proven *efficacy* (effects under ideal circumstances) but may lack *effectiveness* (effects under usual conditions of practice). It may be difficult for clinicians to perform the procedure in the same manner as investigators with special expertise and a standardized protocol. Even in randomized controlled trials, volunteer participants may differ in important respects from the population targeted by clinical preventive measures. The average patient, for example, may be less willing than research volunteers to comply with interventions that lack widespread acceptability. The cost of the procedure and other logistical considerations may make implementation of the recommendation difficult for the health care system without compromising quality or the delivery of other health care services.

Review Process

The Task Force initiated a review process early in the production of this edition by inviting primary care specialty societies and U.S. Public Health Service agencies to appoint liaisons to attend and participate in Task Force meetings. Representatives of the American Academy of Family Physicians, American Academy of Pediatrics, American College of Physicians, and American College of Obstetricians and Gynecologists participated in Task Force discussions and provided expert review by members of their organizations. Similarly, ex officio liaisons of U.S. Public Health Service agencies (Agency for Health Care Policy and Research, Centers for Disease Control and Prevention, National Institutes of Health, etc.) provided access to the expertise of government researchers and databases in examining Task Force documents.

Following this initial review, Task Force recommendations were reviewed by outside content experts in government health agencies, academic medical centers, and medical organizations in the U.S., Canada, Europe, and Australia. More than 700 experts reviewed recommendations included in this report. Recommendations were modified on the basis of reviewer comments if the reviewer identified relevant studies not examined in the report, misinterpretations of findings, or other issues deserving revision within the constraints of the Task Force methodology. The format of this report was designed in consultation with representatives of medical specialty organizations, including the American Medical Association, the American College of Physicians, the American Academy of Family Physicians, the American Academy of Pediatrics, the American College of Obstetricians and Gynecologists, the American College of Preventive Medicine, the American Dental Association, and the American Osteopathic Association.[8]

Conclusion

Recommendations appearing in this report are intended as guidelines, providing clinicians with information on the proven effectiveness of preventive services in published clinical research. Recommendations for or against performing these maneuvers should not be interpreted as standards of care but rather as statements regarding the quality of the supporting scientific evidence. Clinicians with limited time can use this information to help select the preventive services most likely to benefit patients in selected risk categories (see Chapter iii), but no recommendation can take into account all the factors that influence individual clinical decisions in individual patients. Sound clinical decisions should take into account the medical history and priorities of each patient and local conditions and resources, in addition to the available scientific evidence. Departure from these recommendations by clinicians familiar with a patient's individual circumstances is often appropriate.

The draft update of this chapter was prepared for the U.S. Preventive Services Task Force by Steven H. Woolf, MD, MPH.

REFERENCES

1. Battista RN, Fletcher SW. Making recommendations on preventive practices: methodological issues. Am J Prev Med 1988;4(suppl):53–67.
2. Lefebvre RC, Hursey KG, Carleton RA. Labeling of participants in high blood pressure screening programs: implications for blood cholesterol screenings. Arch Intern Med 1988;148:1993–1997.
3. MacDonald LA, Sackett DL, Haynes RB, et al. Labelling in hypertension: a review of the behavioural and psychological consequences. J Chronic Dis 1984;37:933–942.
4. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown, 1985.

5. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials. Baltimore: Williams & Wilkins, 1988.
6. Bailar JC III, Mosteller F, eds. Medical uses of statistics. 2nd ed. Boston: NEJM Books, 1992.
7. Canadian Task Force on the Periodic Health Examination. Canadian guide to clinical preventive health care. Ottawa: Canada Communication Group, 1994.
8. Centers for Disease Control. Chronic disease control activities of medical and dental organizations. MMWR 1988;37:325–328.

Appendix A

Task Force Ratings

The tables of ratings on the following pages were developed for the U.S. Preventive Services Task Force using the methodology adapted from the Canadian Task Force on the Periodic Health Examination[a] and described in Chapter ii. For this edition of the Guide, the Task Force developed ratings for all of the topics examined.

The Task Force graded the *strength of recommendations* for or against preventive interventions as follows.

Strength of Recommendations

A: There is good evidence to support the recommendation that the condition be specifically considered in a periodic health examination.
B: There is fair evidence to support the recommendation that the condition be specifically considered in a periodic health examination.
C: There is insufficient evidence to recommend for or against the inclusion of the condition in a periodic health examination, but recommendations may be made on other grounds.
D: There is fair evidence to support the recommendation that the condition be excluded from consideration in a periodic health examination.
E: There is good evidence to support the recommendation that the condition be excluded from consideration in a periodic health examination.

Determination of the quality of evidence (i.e., "good," "fair," "insufficient") in the strength of recommendations was based on a systematic consideration of three criteria: the burden of suffering from the target condition, the characteristics of the intervention, and the effectiveness of the intervention as demonstrated in published clinical research. Effectiveness of the intervention received special emphasis. In reviewing clinical studies, the Task Force used strict criteria for selecting admissible evidence and placed emphasis on the quality of study designs. In grading the *quality of evidence*, the Task Force gave greater weight to those study designs that, for methodologic reasons, are less subject to bias and inferential error. The following rating system was used.

[a] Canadian Task Force on the Periodic Health Examination. The periodic health examination. Can Med Assoc J 1979;121:1193–1254.

Quality of Evidence

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.

Well-designed and -conducted meta-analyses were also considered, and were graded according to the quality of the studies on which the analyses were based (e.g., Grade I if the meta-analysis pooled properly randomized controlled trials).

An exact correlation does not exist between the strength of the recommendation and the level of evidence, i.e., Level I evidence did not necessarily lead to an "A" grade, nor did an "A" grade require Level I evidence. For example, there may have been evidence of good quality that did not prove that an intervention is effective (e.g., mammography in women under age 50, which received a "C" recommendation). On the other hand, an "A" recommendation was given to screening for cervical cancer with Papanicolaou testing, based on burden of suffering and Level II evidence supporting the effectiveness of the intervention. For many preventive services, there is insufficient evidence to determine whether or not routine intervention will improve clinical outcomes ("C" recommendation). A variety of different circumstances can result in a "C" recommendation: available studies are not adequate to determine effectiveness (e.g., insufficient statistical power, unrepresentative populations, lack of clinically important endpoints, or other important design flaws); high-quality studies have produced conflicting results; evidence of significant benefits is offset by evidence of important harms from intervention; or studies of effectiveness have not been conducted. As a result, lack of evidence of effectiveness does not constitute evidence of ineffectiveness. Chapter ii provides further information about the methodology used to develop the body of this report.

Table 1.
Screening for Asymptomatic Coronary Artery Disease

Intervention	Level of Evidence	Strength of Recommendation
Routine resting, ambulatory, or exercise electrocardiography in middle-aged or older persons	II-2	C
Routine resting electrocardiography in healthy children, adolescents, or young adults, including those undergoing pre-participation sports physicals	III	D

Table 2.
Screening for High Blood Cholesterol and Other Lipid Abnormalities

Intervention	Level of Evidence	Strength of Recommendation
Routine measurement of total serum or blood cholesterol in		
Men aged 35–65 yr	I, II-2	B
Women aged 45–65 yr	II-2	B
Persons aged > 65 yr	II-2	C
Children, adolescents, young adults	II-2	C
Routine measurement of HDL-C	II-2, III	C
Routine measurement of triglycerides	II-2	C

Table 3.
Screening for Hypertension

Intervention	Level of Evidence	Strength of Recommendation
Periodic blood pressure measurement in persons aged ≥ 21 yr	I	A
Measurement of blood pressure in children and adolescents during office visits	II-2, II-3, III	B

Table 4.
Screening for Asymptomatic Carotid Artery Stenosis

Intervention	Level of Evidence	Strength of Recommendation
Routine carotid ultrasound or auscultation for carotid bruits in older persons	I, II-2	C