Methods for Prioritizing Clinical Preventive Services: Technical Report Prepared for the National Commission on Prevention Priorities

DRAFT 0.1; last updated May 15, 2006

This DRAFT is under-revision.

Please return to this website to check for an updated report.

Questions regarding methods not answered in this draft may be addressed to the authors.

Prepared by

Michael V. Maciosek, PhD*

Ashley B. Coffield, MPA**

Nichol M. Edwards, MS*

Thomas Flottemesch, PhD*

Michael J. Goodman, PhD*

Leif I. Solberg, MD*

*HealthPartners Research Foundation
8100 34th Ave S
PO Box 1524, MS 21111R
Minneapolis MN 55440-1524

** Partnership for Prevention
1015 18th Street NW, Suite 200
Washington, DC 20036

This work was supported by the Centers for Disease Control and Prevention (CDC) and the Agency for Healthcare Research and Quality (AHRQ).

1. INTRODUCTION

In 2001, we published the first known study to apply structured analyses in a consistent fashion across a full set of recommended clinical preventive services to determine their importance to the U.S. population.[1] For decision-makers, knowledge that a clinical preventive service is effective is not sufficient to set priorities for delivery of preventive care. Resources (including clinician and patient time) are limited, and services differ in their potential health impact and costs. We presented a priority ranking of 30 clinical preventive services recommended by the 2nd U.S. Preventive Services Task Force[2] based on their relative value to the U.S. population.

The new ranking reflects updated methods. The 2001 methods [3] were well-suited for an initial effort to inform priorities setting among clinical preventive services and have been proposed for use in other endeavors.[4;5] However, areas for improvement remained. The updated methods take advantage of what we learned about the data needs and availability in the initial ranking exercise, and they address several constructive criticisms of the initial methods. The focus of our methods improvement is the use of a more systematic and transparent process for searching, tracking, and abstracting literature. This technical report describes all of our methods including these and other improvements.

The first two steps in developing a ranking of clinical preventive service are to 1) identify preventive services to be included in the ranking and 2) choose the measures on which the ranking is based. The National Commission on Prevention Priorities, a 30-member panel chaired by David Satcher and consisting of researchers, health plan executives, employers, and state and federal health officials, made these decisions.

The Commission chose to include select services recommended by the USPSTF and the Advisory Committee on Immunization Practices (ACIP) through December 2004. Specifically, the ranking includes services recommended by the USPSTF for the general population, vaccines recommended by the Advisory Committee on Immunization Practices for the general population, and services recommended by the USPSTF for persons at high risk for cardiovascular disease. These criteria exclude clinical preventive services for which the USPSTF recommends against providing or found insufficient evidence to recommend, clinical preventive services not reviewed by the USPSTF, and all community (or population-based) preventive services and programs.

The Commission chose to base the ranking on the measures of our previous effort: **clinically preventable burden (CPB), which measures services' health impact, and cost-effectiveness (CE), which measures services' value**. These measures were originally selected by the Committee on Clinical Preventive Service Priorities, which guided the 2001 study, and were adopted by the Commission. We refer readers to our prior discussions of the choice of CPB and CE to rank services.[1;3]

After these decisions have been made, the key challenge of setting priorities among clinical preventive services is to derive consistent estimates of services' value using disparate data. A valid ranking of preventive services requires each service's to be evaluated on the same basis.

We focus this report on the steps taken to insure valid measurement of CPB and CE. Our methods are designed to make our estimates of the magnitude of health benefit (CPB) and cost-effectiveness comparable across services through a systematic search and summary of available data for each service. Our methods differ from those used in other systematic literature reviews because our goal is to quantitatively compare the value of multiple services. If instead our goal was to systematically assess the literature to determine whether or not a single intervention was effective, or to determine if a relationship between a risk factor and a disease exists, we could readily develop and apply evidence inclusion criteria and exclusion criteria that suited the specific body of literature examined. However, our goal is to aid decision-makers by producing comparable estimates of CPB and CE from available data for multiple clinical preventive services. Doing so requires the evaluation of all important variables for each service, not just those for which a substantial body of evidence exists. By necessity, some of the included data would not meet the inclusion criteria of many systematic reviews.

At the same time, the methods must help to identify the best available data when multiple estimates were available, where the best data are those that produce accurate estimates of the health benefit and cost-effectiveness of the service in usual practice for the general US population. Thus, for example, we need to use estimates that produce the most accurate point estimate for effectiveness, and not necessarily just randomized control trials that provide the best indication of whether or not the a service is effective.

Due to the judgments necessary to identify the best estimates we do not believe it is possible to design a practicable set of methods for ranking a diverse set of preventive services that allows reproducibility of results. The methods must allow for judgments to accommodate the diversity of data needs and disparity in data availability among the services. Judgments are necessary when identifying the most appropriate data within articles, determining when an estimate from a marginally applicable study adds to or detracts from the estimate produced by a body of evidence, making decisions about secondary outcomes or treatment options which are too insignificant to the value of the service to justify extensive literature review, and others as described below.

We strove to develop a system in which judgments have a limited impact on final results, such that judgments would not lead to a change in the final score for a service by more than one point in the scoring system (see Section 5). In most cases, the available data do not allow precision beyond +/- one point in the scoring system. Therefore, the goal of our methods is to keep the judgments made by the study team within the margin of error inherent in the available data. To improve transparency so that others can see where our decisions may have affected an estimate for CPB or CE, these decisions are documented in each service's technical report available online at www.prevent.org/ncpp. The technical reports will be made available as background articles for each service are released. At the time of this version of the methods technical report, technical reports are available for breast cancer screening, cervical cancer screening, colorectal cancer screening, influenza immunizations, and screening for tobacco use and brief counseling.

Based on experience from the 2001 study, we designed methods that balance consistency in principles with flexibility in application in order to accommodate differing characteristics of services. It is impractical to enumerate the necessary variations between principles and application for services for all services in a single document. Instead, these details will be available in technical reports for each service. In this report we explain the principles to improve the consistency across services. We report issues that were encountered among multiple services and the decisions we made to address them. Issues that were specific to a single service are discussed in the technical report of that service rather than in this methods summary.

2 DATA COLLECTION AND SUMMARY

Thousands of data points underlie the priority ranking, including the incidence of 20 diseases attributable to tobacco use, duration of illness for all diseases and injuries addressed by preventive services, components of effectiveness such as sensitivity of screening and adherence with recommended behavior change, and costs ranging from preventive service delivery costs to medical costs of prevented treatments. Tens of thousands of journal articles have been written either about the preventive services themselves or about the burden of disease and costs of the conditions addressed by the services. In addition to these, book chapters, government reports, public use data sets, unpublished academic papers, and non-profit association papers and internet publications provide potentially useful data. The key challenge in using this information to produce a valid ranking to devise a data search strategy that identifies the best estimates while minimizing the amount of resources spent identify, gathering, and evaluating estimates that ultimately add little precision to the ranking. We accomplished this through using two tools:

standardized search levels to ensure consistent evidence search strategies; and

systematic literature abstraction to ensure consistent evaluation of the literature.

2.1 Evidence Search Strategies

Our evidence gathering methods were developed to meet three objectives: 1) provide transparency of our search strategy; 2) collect the vast majority of relevant evidence; and 3) minimize study resources spent searching, collecting and analyzing evidence that does not improve the reliability of the ranking.

2.1.1 Standardized Search Strategies

To meet these objectives we developed the 2 Standardized Search Strategies shown in Figures 1A and 1B. The first set of strategies are for effectiveness and cost-effectiveness data and the second set is for burden of disease and cost data. Each set of strategies defines 4 search levels where the first level includes the most current literature and data sources and each subsequent level extends to less current sources and sources less likely to yield useful data. For each piece of data, we first employ the strategies defined by search level 1. We then evaluate the quantity and utility of data found (see Section 2.3) and determine if extending the search to the next level is likely to improve our estimate of either CPB or CE.

Search level 1 covers the most current literature and important articles from older literature. Therefore, the core of all available evidence is captured in Search Level 1 and in many cases it is not necessary to move to Search Levels 2, 3, or 4. However, for topics on which data are scarce we move to subsequent levels to get the best possible estimate from limited information.

The overarching study goal was to make it easier for decision makers to use the best-available evidence, even when the best available evidence is limited in quantity, quality or generalizability. In comparing a wide range of preventive services, the quantity and quality of the data vary greatly between preventive services and frequently among multiple health conditions addressed by a single preventive service. The Standardized Search Strategies directed the study team toward more extensive searches when necessary, while not expending resources on extensive searches when the best data were easily identified.

At the same time the Standardized Search Strategies improve transparency by allowing us to report where we terminated our search for evidence. We report the search levels used in each service's technical report, and we keep a record of the specific search strategies (eg. Medline keywords and limits).

The principal limitation of the search strategies is that the study team needed to make judgments about whether to base estimates on data found using search level 1 or move to subsequent search levels. We erred toward a conservative approach by exploring the next search level when there was reasonable doubt that the prior search level(s) have identified the best available data.

2.1.2 Use of systematic reviews

We typically do not use the results of systematic reviews of intervention effectiveness directly in our estimates. Systematic reviews rarely include the same estimates from studies that we need to best estimate the magnitude of effect size, and they rarely report the detail from each included study that is needed to determine how appropriate the results are as a measure of effect size in usual practice. For example, intention-to-treat estimates from randomized controlled trials reflect non-adherence of the selected and self-selected participants in the trials that is likely to be substantially lower than non-adherence in all primary care patients. Similarly, some non-randomized studies have design and analysis limitations that call into question the accuracy of their estimated effect size. Therefore, we found it necessary to evaluate each individual study ourselves and calculate our own effect size to produce an estimate of CPB that is accurate and consistent with estimates for other services in the ranking. This also allowed us to include more recent studies that are not part of published systematic reviews, and use different inclusion and exclusion criteria as necessary to produce the best possible estimate of effect size. As a result, our estimated effect size of a preventive service is typically somewhat different than that reported in any available systematic review.

We did use systematic reviews to reduce study resources devoted to literature searches. When recent comprehensive literature reviews were available, we used the literature review as a starting place for our review. We then supplemented the review by searching for studies published since the completion of the systematic review. In addition, we searched for studies that were explicitly excluded by the systematic review as appropriate. For example, if a systematic review limited its scope to randomized control trials, we performed a supplemental search for observational studies that may provide better estimates of the effect size in usual practice.

We also used reviews, including less comprehensive reviews, to check the results of our searches. If reviews identified one or more important articles that our literature search strategy failed to identify, we examined our search strategy to determine what changes were needed to capture the missed article(s) and others like it.

2.1.3 Searches for components of effectiveness to create indirect estimates of effectiveness

When available, we first abstracted effectiveness articles (see literature abstraction process below) that provided estimates of the effectiveness of the preventive service in reducing the final health outcomes (disease cases, sequelae and mortality). When we found insufficient direct evidence on effectiveness of the preventive service in preventing all important disease and mortality, we expanded our search to include other links within the analytic framework (such as adherence, screening sensitivity, and treatment effectiveness) in order to calculate an indirect estimate of effectiveness using the individual pieces of data. The effectiveness abstraction form prompts reviewers to record all reported data on all key components of effectiveness in addition to any estimates of reduction in disease or mortality. From these we created an indirect estimate of effectiveness using the following calculation:

% Effectiveness = % adherence with screen × % cases detected by screen × % adherence with follow-up × % effectiveness of treatment in reducing mortality.

In practice, the most direct route to estimating effectiveness for screening usually involved at least two links, adherence with invitations for screening and effectiveness of screening in reducing mortality (effectiveness = % initial adherence x % reduction in mortality with screening). For counseling services, it was typically necessary to calculate an indirect estimate of effectiveness using the effectiveness of counseling at producing a long-term behavior change and the effectiveness of behavior change at reducing disease or injury risk.

2.2 Evaluation of Literature

After potential literature is identified, abstracts are reviewed to determine whether the study may have data relevant to our analyses. All studies are obtained that we are unable to exclude based upon the abstract. Basic data from the obtained studies, such as intervention description, study sample size and characteristics, and outcome measurements, are entered into a table and a subset of studies appearing to meet inclusion criteria for that particular service are reviewed in more detail.

2.2.1 Literature Abstraction

As do other systematic reviews (Briss; Carande-Kulis; USPSTF?; Cocharane?),[6-8] we used literature abstraction forms to ensure consistent evaluation of the literature. The study abstraction forms are composed in Microsoft Access. Reviewers complete the forms electronically, and the data are automatically stored in an Access database. Versions of the effectiveness and cost-effectiveness abstraction forms (formatted for printing) will be made available in the next version of this technical report. In the meantime, interested readers may obtain the electronic form in Microsoft Access file format and instructions from the authors.

The abstraction forms differ from those used by others is some basic respects. First, the abstraction are designed to be flexible in terms of data entry in order to allow the same form to be used across services. The data entry forms have more text boxes for open-ended abstractor comments and fewer check boxes than a form that is designed for a single service. This choice was made to avoid to avoid spending study resources developing and testing abstraction forms for each service.

Second, the abstraction forms include an evaluation of 'utility of study'. Reviewers evaluate the usefulness of the study for the purpose of completing a ranking of clinical preventive services. There are 6 utility criteria in the effectiveness abstraction, and 5 in the cost-effectiveness abstraction form as shown in Figure 2. Reviewers rate each study on a simple 3-point scale for each criterion and enter a brief explanation for their criteria scores. A total score is calculated as an un-weighted average of the criteria scores.

There is an important distinction between what we call study 'utility' and study quality. Utility is the usefulness of the article in providing data point(s) that can be used to generate a reliable estimate of the CPB or CE of a preventive service according to the definitions we apply consistently across services (see Sections 3 and 4). Some markers of study quality are important for this determination, including those related to internal and external validity of the study.[9-13] However, studies are often designed to inform decisions that are somewhat different than those of our study. Therefore, higher-quality studies may be of lower utility for our purposes, and lower-quality studies may occasionally receive a relatively high utility score if they provide a key piece of data that is not greatly affected by the study's quality short-comings.

The utility scores were not used in a quantitative manner in estimating CPB or CE or in calculating the ranking. Instead they were used as to alert the study team to potentially biased or otherwise inappropriate estimates. Lower utility scores warn the study team to carefully review the study before using the study results to estimate CPB or CE. In addition to the criteria rating shown in Figure 2, each reviewer has the opportunity to flag the study with a 'fatal flaw' indicating that the reviewer believes the study results are clearly not suitable for use in estimating CPB or CE due to an important quality or other utility concern.

To limit the impact of reviewer judgments and to reduce errors of omission and interpretation, two reviewers abstracted each effectiveness and cost-effectiveness article. To reduce non-subjective duplicative data entry work, the second reviewer received the form with the results (effect size or CE ratio) from the first reviewer filled-in. The two reviews were otherwise independent. The second reviewer completed all sections on study design, study population and analysis; reviewed the data entry of the results for accuracy and omissions; and completed an independent assessment of study utility.

2.2.2 Adjudication

One member of the study team began an adjudication process by first comparing the completed abstraction from both reviewers. During an adjudication meeting, reviewers discussed differences in their entries and come to an agreement on a single final 'adjudicated abstraction form'. The reviewers discuss their utility scores as part of the adjudication. Each reviewer was permitted to change a score based upon discussions during the adjudication meeting, but reviewers do not attempt to settle on a single utility score for each criteria. Instead, the average of the two reviewers' scores was placed in a final adjudicated abstraction form that reflected the results of both reviews and the adjudication process.

2.2.3 Inclusion and Exclusion Criteria

Stringent inclusion and exclusion criteria were generally not applied a priori. It was necessary to examine a wide variety of literature to determine the best available estimates for each data point. In cases where large amounts of high quality studies were available, it was possible to obtain the best available estimates by restricting the literature that was ultimately included in the estimates. In other cases, very limited literature was available and strict inclusion criteria would eliminate most or all available evidence. The lack of strict inclusion and exclusion criteria result in some uncertain data points, but allowed us to meet the study's objective of providing decision makers with estimates based upon the based available data, even if that data is limited.

A priori inclusion and exclusion criteria that were to limited the set of studies abstracted to those meeting the gender and age-group recommendations for the service of the USPSTF or ACIP, and to limit the technologies used in the preventive services and follow-up to those recognized by the USPSTF as evidence-based. For example we did not include literature on computed tomography (CT) colography for colorectal cancer screening because the USPSTF found insufficient evidence to support its use for screening.[14] Nor did we include literature on diet or exercise counseling following cholesterol screening because the USPSTF based its recommendation on 'good evidence that lipid-lower drug therapy substantially decreases the incidence of coronary heart disease' and found no direct evidence in the literature that dietary or exercise intervention delivered in primary care prevent coronary heart disease events.[15;16]

Two other exclusion criteria were applied to the analysis of more than one service. When a body of literature contained a fair amount of large and small studies, we generally limited the literature we abstracted to studies with at least 25 participants in each study arm. For studies requiring behavioral change such as smoking cessation, modifying alcohol use, increasing physical activity or modifying diet, we limited studies included to those that measured outcomes at least 6 to 12 months (depending on the available literature) post intervention.

Other exclusion criteria were specific to the characteristics of individual preventive services and the available literature. These are described in each service's technical report. In making judgments about whether or not to include service components, health outcomes and costs in the estimates for each service we balanced their potential quantitative importance and the quality of available data. Guidelines for conducting cost-effectiveness studies recognize the need to make a priori choices about measuring lesser outcomes(refs). For model components where only poor quality data available and the model component is potentially important to the final estimate we determined and included the best possible estimate in order to produce the most accurate estimate possible. For model components for which there was no potential for the component to be quantitatively important, we made judgments on a case-by-case basis as to whether including poor quality estimates was likely to improve or reduce the precision of the estimate. Sensitivity analysis on preliminary results were used to gage the potential quantitative importance.

We excluded potential outcomes of a preventive service for which evidence of an impact are lacking. For example, due to conflicting evidence we excluded the theoretical benefits of most cancer reduction from intensive dietary counseling. When missing, weak or conflicting evidence creates uncertainty as to whether or not there is a connection between a preventive intervention and a health event, it is difficult to quantify what that size of the benefit may be.

2.2.4 Summarizing evidence

To assemble the body of evidence on effectiveness and cost-effectiveness, key data from the adjudicated abstraction databases were placed in spreadsheets. We used the mean value across the collected estimates with two exceptions: 1) when the number of estimates is small and it appears possible that the estimated mean may be pulled away from a 'true' mean by one or two studies, we use the median value; and 2) when we have a large number of estimates, and some are clearly better than others for our purposes as indicated by the 'utility scores,' we use the mean of the better articles. Due to study quality or applicability issues, some data are at least as likely to distort our effectiveness or CE estimates as they are to improve them. For both effectiveness and cost-effectiveness we have noticed that such articles generally receive a utility score of less than 2 (on a scale of 1 to 3).

We did not limit included studies to those with statistically significant results. To do would have risked biasing summary estimates upwards as small studied with large effect sizes would be included while equally small studies with small effect sizes could have been excluded.

3 BASIC MODEL STRUCTURE

The approach to estimating CPB and CE differed from standard Markov models in three ways. First, we calculated CPB and CE at the population average rather than at the individual level using a hypothetical population with a defined distribution of characteristics and events. Second, using population averages, we calculated health benefits as the cumulative, long term, benefits over rather than through year-by-year transition probabilities. For example, we based the benefits of cholesterol screening upon average long-term adherence with therapy and the long-term efficacy of therapy in preventing heart disease rather than through yearly probabilities of continued adherence and the projected benefits of one-year's adherence on future heart disease events and associated costs. Third, we based the estimates on the average experience of patients as reported in the effectiveness literature rather than models of each specific clinical pathway that different patients may experience and the probability of each pathway. Readers who wish to see exactly how calculations were performed may consult the technical reports for the services.

Some preventive services have multiple options, either for the initial preventive service, or for follow-up treatments. For example, there are four effective options for colorectal cancer screening, and multiple pharmaceutical interventions to treat hypertension following initial screen. Each option is likely to have a different level of effectiveness, and in practice, each option will be chosen by some portion of the population.

In order to estimate CPB in such cases, we calculate a weighted average of effectiveness where the weights are the percent of those who choose each option among those currently receiving the preventive service. This is identical to dividing our estimate of the total burden of disease according to the same percentages and applying individual estimates of effectiveness.

To use study resources most effectively, options used by a very small portion of the population and are similar in effectiveness to other options are sometimes excluded from the estimates. In effect, this adds a small amount of error by substituting the average effect the included options for the portion of patients for which the excluded option is chosen. When use of the excluded option is small or effectiveness is not substantially different from the included options, the impact on the CPB estimate is small while the time savings from formal literature abstraction can be substantial.

4 CONSISTENT ESTIMATION OF CPB

After data were collected and summarized, we used it to estimate CPB and CE using methods that ensure valid comparisons across services. This section summarizes the principles that promoted consistency in measuring CPB. In practice, these principles also provided direction in the data collection processes. Our linear presentation of these methods understates their interconnections. Readers will be able to readily infer from the description of data collection above how the principles that promote consistency in CPB described below inform data collection. Therefore, the following discussion focuses on using the data once they are in-hand.

4.1 Conceptual CPB formula

Conceptually, CPB is defined as the burden of disease in the target population multiplied by the percent reduction in the burden that can be achieved by the preventive service.

CPB = burden of disease x % effectiveness of the service

Burden of disease has two primary components: morbidity and mortality. Effectiveness has several components that vary from one service to the next.

4.2 Formal Definition of CPB

Given the varied nature of each service and its associated literature, each service presents unique situations that require judgments regarding which data to include and how available data should be used to build the best possible model for CPB. To maintain consistency across services while these judgments are made, we devised a formal definition of CPB that embodied five major principals. Comparability across services was guided by referring to this definition when making decisions for any particular service.

> **Clinically Preventable Burden** is the total quality adjusted years of life that could be gained in typical practice if the clinical preventive service were offered at recommended intervals to a U.S. birth cohort of 4,000,0000 over the years of life the service is recommended.

4.3 Measuring Total Burden of Disease

The definition of CPB includes the 'total quality adjusted year of life' obtainable by the service in typical practice. There are two major principals underlying this part of the definition. First, the burden of disease that underlies the calculations of health benefits must include both morbidity and mortality. Second, that burden should include both the burden that exists today and the burden that is already being prevented by delivery of the service.

4.3.1 Measuring QALYs: Years of life lost, duration of illness, and quality of life reduction

For clarity it is important to note that our methodology is *not* based on a concept of "burden of disease of the service". Although each service targets particular diseases and sequelae, there is no natural definition of 'burden of the service' that can be consistently applied to all services. For example, the burden of disease associated with counseling to promote folic acid supplementation could be defined as all births with neural tube defects or neural tube defects attributable to insufficient folic acid. The burden of disease associated with cholesterol screening could be defined as all cardiovascular disease (CVD), all CVD caused by high cholesterol, or all CVD (regardless of cause) in people with high cholesterol. In each of these cases, the available effectiveness estimate dictates which burden of disease definition is relevant for the analysis. To create an accurate estimate of CPB, the burden data must capture all important sequelae that are impacted by the clinical preventive service and the effectiveness estimate must measure the same health outcomes in populations with similar risks. It is not valid, for example, to apply an estimate of the percent effectiveness of hypertension screening in reducing myocardial infarctions in persons with hypertension to all myocardial infarctions occurring in the US population.

No matter how burden is defined for each service, it has two potential components, morbidity and mortality, that must be combined into a single measure to derive a measure of CPB that can compared across services. The measure we have chosen is generically known as the Quality Adjusted Life Years (QALYs)(refs). As the name implies, it is a measure of years of life that are adjusted for quality of life.

Mortality is encompassed in QALYs by measuring years of life. Years of life are measured as the life expectancy from any given age. Thus, if a coronary heart disease death occurs at age 75, the years of life lost are equivalent to the life expectancy at age 75.

Morbidity is measured as a reduction in quality of life. For example a person living at 70% of optimal health for one year due to a chronic disease would have a quality of life reduction of .30 years. In this manner, both fatalities and morbidities are expressed in years of life equivalents.

Measuring morbidity in this way requires 3 components: incidence of disease or disability, duration of disease, and the extent of the reduction in quality of life over the duration of disease. A small number of tables of disease- and condition-specific QALY weights (refs) and disability weights (refs) have been published, and each uses different methods for determining the weights. In these tables, regardless of method, all but the most severe and least severe conditions fall in the range of approximately 0.5 to 0.9 for short-duration illness or injury and 0.7 to 0.9 for chronic conditions. No published table of condition-specific weights derived using a single method includes all of the health conditions addressed by the set of clinical preventive services included in this analysis. For our base-case estimates, we adopted the midpoints of these ranges: 0.7 for short-term conditions and 0.8 for chronic conditions.

Compared to perfect health valued at 1.0, an illness or injury resulting in a QALY weight of 0.7 would indicate a quality of life lost of 0.3 QALYs per year (1.0 – 0.7 = 0.3). Most weighting systems find that persons without existing medical conditions have a quality of life of less than 1.0, usually about 0.9. A more precise measure of the quality of life lost to a medical condition with a weight of 0.7 would therefore be 0.2. However, rather than re-scaling all years of life lost and all quality of life lost to a 0.9 scale, we evaluated years of life lost at 1.0 and estimated quality of life lost to morbidity by subtracting the QALY weight from 1.0. This approximation slightly overstates burden of disease across most services and therefore has little or no impact on ordering of service by CPB.

Is important to note that this assumption also impacts the tabulation of years of life lost because all years are assumed to be lived in perfect health of 1.0. Improving accuracy by removing the comparison to perfect health would have required data on the average health related quality of life by age group, gender and risk factors. For example, to most accurately compare the CPB of breast cancer screening and influenza immunizations, we would have needed data on the baseline quality of life for women who will develop breast cancer and older adults who will suffer the consequences of influenza infections. As with quality of life data for particular conditions, such estimates are not available from a single method, if available at all. It is not clear that any assumptions we could have made for regarding the relative quality of life in the absence of the condition of interest for different populations would produce more accurate estimates than our simple, transparent assumption applied to all services.

Exceptions to the QALY weights were made for those conditions shown clearly by the available information to be either clearly more or less severe than for other conditions that fall within the ranges of .5 to .9 for short-duration illness and .7 to .9 for chronic conditions. Base-case estimates were assigned based on review of estimates from the published tables(7, 8, 9). Other exceptions were made when a service's estimates of CPB and/or CE were particularly sensitive to the estimate of QALYs lost. These services primarily or exclusively addressed morbidity instead of mortality. For conditions addressed by these services, published scales were reviewed to identify more precise base-case estimates.

Duration of illness or injury was also an important consideration and is logistically related to the QALY weight. In general, the severity of a condition will change over time. Ideally, we would have employed measures of quality of life for each day for which is a condition is suffered and we would measure total quality of life lost by summing each day's quality lost over the duration if illness. In practice, we generally found a single measure of quality of life lost for a condition which represents, to various degrees, the average quality of life over the duration of illness. The duration of illness to which the average quality of life is applicable is usually not specified when QALY weights are reported.

For many acute conditions, we relied upon medical encounters (e.g., hospital inpatient stays, emergency department visits, and other outpatient visits) to approximate incidence. The average severity and duration of disease is likely to vary by type of medical encounter. When only estimates of emergency department visits or hospital inpatient stays were available, we presumed that only the more severe cases were represented and assigned a longer duration of illness than if outpatient estimates had been available. Thus, we varied the average duration of illness according to the type of incidence data available, rather than varying our standard QALY weight.

Actual data on the duration of morbidity for acute conditions were rarely available. Estimates of restricted-activity days (including, but not limited to days spent in bed, days lost from work, and days lost from school) have been tabulated from the National Health Interview Survey (ref). For other acute conditions, we assigned duration based on the perceived severity and likelihood of extended disability relative to the conditions for which restricted-activity days have been tabulated.

For chronic conditions, it is generally possible to identify more accurate estimates of incidence and duration. Individual studies frequently provide estimates of annual incidence, though occasionally only point prevalence estimates could be identified in individual studies. For many chronic conditions, we used incidence rates and average duration estimates for Established

Market Economies from the Global Burden of Disease Study.[17;18] For a few conditions, duration was assigned based on life expectancy at the estimated average age of onset.

The number of approximations necessary to estimate morbidity prevented across a broad range of clinical preventive services may raise concerns about the accuracy of estimates of CPB. However, mortality incidence is the most important variable for most services. For two services (counseling on calcium supplementation and screening for osteoporosis), we used a published tabulation of QALYs lost per fracture and there for the methods outlined above do not apply. In 18 of the remaining 23 services, morbidity is 25% or less of CPB. Only for the remaining 5 services does the choice of QALY weight and duration have a potential impact on the CPB score, and in these cases, the CPB score is extremely unlikely to change by more than one as a result of changes to QALY weights or duration of morbidity.

CE estimates are even less sensitive to changes in the morbidity components. In some cases, CE may be sensitive to counts of hospitalizations underlying the morbidity calculations. However most estimates of hospitalization are derived from a national survey and therefore are only limited by the accuracy of diagnostic coding.

4.3.2 Accounting for current delivery of the service

The amount of disease that is observed today is influenced by how frequently preventive services that address each particular disease have been delivered, and how effective those services are. Services that are frequently delivered and are very effective, such as childhood immunizations, have greatly reduced the burden of disease we observe today. If we were to calculate CPB based upon the observed burden, the value of these services would be greatly understated. We base our estimates on the 'total burden' of disease that would be observed in the absence of the preventive service.

We used three methods for estimating burden of disease in the absence of the service. For some services, historical incidence rates – prior to widespread provision of the preventive service – provide the best estimates of total burden. This is true for many childhood vaccinations for which wide-spread provision of the vaccines is providing some protection to all individuals through herd immunity. For other services, current incidence rates in persons not receiving the service can be used, if the population in which these incidence rates are observed are representative of the US population. In this method, incidence rates from these populations are generalized to the entire US population with the assumption that those currently receiving the preventive service would, in the absence of the service, be at the same risk as those who did not receive the service.

For some services, the only, or best available data are national incidence rates that reflect both those who did and did not receive the service. To use these data we estimated total burden by adjusting currently observed burden for current delivery rates and effectiveness of the service. When total burden is estimated in this way, it was calculated as:

$BD_{ZDR} = BD_{CDR} + [1 - (CDR \text{ x } Eff)]$

which is derived from substituting equation (2) below into equation (1) and algebraic manipulation:

(1) $BD_{ZDR} = BD_{CDR} + BD_{PCDR}$

where,

BD_{ZDR} = projected Burden of Disease with zero delivery rate of the clinical preventive service;

BD_{CDR} = currently observed Burden of Disease at the current delivery rate of the preventive service; and

BD_{PCDR} = Burden of Disease prevented at the current delivery rate

(2) $BD_{PCDR} = BD_{ZDR} \times CDR \times Eff$

where CDR and Eff = Current Delivery Rate and Effectiveness, respectively, of the preventive service.

In the prevention priorities project our conceptual definition of 'delivery' is the act of a clinician offering the preventive service to a patient, and 'adherence' is the patient's actions in accepting a service and subsequent follow-up (see section 4.5 below). In practice, available data on delivery rates often reflect the percent who receive the service, without distinction between whether individuals were never offered the service or were offered and refused. Therefore, care was necessary when using delivery rates in order to insure we were consistent with our conceptual definition of 'delivery rates' while not inappropriately applying the available data. For example, when adjusting colorectal cancer mortality rates for current delivery, the available delivery rate data indicate those who received screening, not those who were offered screening. Therefore, in the formula above we use a modified estimate of effectiveness that excludes incomplete adherence with acceptance of the screen.

The adjustments for current delivery rates and effectiveness were limited exclusively to the services being analyzed. For screening for colorectal cancer, for example, we only adjusted the CPB estimate for the absence of screening, not for the absence of tobacco cessation counseling.

For some services, national incidence rates were used without adjustment because low delivery rates combined with low effectiveness have had little impact on burden of disease. We judged that the precision of the estimate was unlikely to be improved if the adjustment calculations revealed the impact on incidence was less than one percent.

4.4 Effectiveness in typical practice: Effectiveness, efficacy, and patient adherence

For all preventive services, careful consideration of the extent to which incomplete adherence reduces the practical benefit of the service was necessary. Conceptually, we distinguished between patient adherence and provider delivery of services. 'Delivery' is the act of a clinician offering the preventive service to a patient, and 'adherence' is the patient's actions in accepting a service and subsequent follow-up.

The importance of adherence varies greatly from one service to the next as the type and magnitude of non-adherence varies. The types of adherence that most frequently impact the CPB and CE of preventive services include acceptance of the preventive service when offered by a clinician, medical follow-up necessary to achieve health benefits (making and keeping referral appointments; completing follow-up diagnostics and treatment plans; beginning and continuing prescription drug use; purchasing and using recommended medical equipment), and behavior modification. Services for which long-term behavior change is needed to achieve the desired health outcomes are particularly sensitive to incomplete adherence. However, adherence is also important to other preventive services. For individuals to receive a health benefit from cancer screening services, they must accept the offer to be screened, comply with follow-up for diagnostics of positive test results, and usually complete a separate treatment plan.

The extent to which adherence is reflected in the effectiveness estimate abstracted from the literature varies from study to study and from one preventive service to the next. Estimates of effectiveness in clinical trials with disease or mortality endpoints typically involve incomplete adherence with treatment. However, because the population includes study volunteers who are likely to receive closer follow-up than would be found in usual care, adherence with treatment may be overstated. In addition, the magnitude of adherence reported depends on how the data were analyzed. Estimates based on intention to treat analysis generally reflect more complete estimates of adherence than estimates that include only those for whom follow-up measures were available.

Therefore, it is often necessary to supplement effectiveness estimates with additional data on adherence. Data on adherence have several sources. The effectiveness abstraction form includes data fields for recording adherence. When abstracting results of effectiveness articles, reviewers also record any reported adherence. However the adherence data contained in many effectiveness articles is of limited generalizability. For example, randomized control trials have study members who have previously agreed to participate in the study. Also, study members may be selected based on participants' statements of willingness to complete study protocols.

In addition to patient adherence, the difference between effectiveness in usual practice and efficacy includes clinician adherence with recommended protocols, frequency of patient contact and monitoring, and selection of patients into clinical trials. We focus on patient adherence because it is quantitatively important and, as a practical matter, adherence is the only dimension which data exist to help measure the difference between usual and typical practice.

Effect size estimates from randomized control trials are usually reported using intention to treat analysis. Therefore they usually reflect some level of patient non-adherence. Typically, however, careful selection of trial participants and intensive monitoring result in better adherence than would be seen in typical practice. Therefore, as data allowed, we calculated a effect sizes more representative of typical practice by 'taking-out' non-adherence in the trial and adding in non-adherence from obtained from observational studies. In effect, this is done with the simple calculation: effectiveness = (trial effect size / trial adherence) × observational adherence.

This simple calculation provides a better estimate of effectiveness in typical practice, but is limited in two important ways. First, individuals who are adherence in either randomized trials may have different risk profiles than individuals who adhere. As a result, the adjustment may over- or understate the effect size, depending on whether those who adhere were at higher or lower baseline risk for the trial end-point. Second, adherence is usually available only for the major component of the trial, such as adherence with screening in the case of screening trial, or adherence with pharmacotherapy in the case of drug trial. Other components of adherence such as follow-up with diagnostic testing and other therapies are typically not reported in trials, and are also less likely to be available for observational studies. As a result, our estimates of effectiveness that are based on randomized control trials often reflect higher levels of adherence for some components of adherence that can be obtained in typical practice.

On the other hand, the effect sizes based on clinical trial data typically reflect contamination of the control group which will cause the effect size to be understated when measuring effect size relative to no intervention. Our data abstraction form captures any reported data on contamination of the control group, but we found that this data is infrequently reported for preventive services literature we reviewed.

Because adjusted estimates from randomized control trials are imperfect, we cannot be sure that they represent better estimates of effect size than those reported from observational studies. Therefore, we do not as a rule, exclude observational studies from our literature review. Typically, we obtain an average effect size from randomized trials (with adherence adjustment), retrospective case-control studies, and retrospective observational studies. Other study designs are included as available such as prospective observational studies or quasi-randomized trials. Y Weaker study designs, such as time-series are used when no other data are available, as is the case for cervical cancer screening, or, as in the case of childhood immunizations, they provide the best available data because they reflect herd immunity. Observational studies may be excluded on a case-by-case basis if self-selection combined with other study short-comings clearly compromise the potential accuracy of the study effect size.

4.5 Target Population, Time Frame, and Delivery Interval

The definition of CPB encompasses the broad definition of the target population that is applied in all analyses. The target population is a U.S. birth cohort of 4,000,000, to whom the service is offered at over the age range recommended by the USPSTF or ACIP. The target population is limited by gender for services recommended only for women (breast cancer screening, cervical cancer screening, Chlamydia screening, osteoporosis screening, prescription of calcium, and prescription of folic acid).

The birth-cohort approach yields an incidence-based estimate of CPB and it guides comparable estimate of CPB and CE across services in two ways. First, by using a birth cohort of 4,000,000 for all services, we remove variability in population size that would occur if we had used a cross-sectional (or point prevalence) approach services targeted at older age groups would have had relatively lower CPB estimates because other birth cohorts were substantially smaller than recent birth cohorts. Over the last several decades, birth cohorts in the US have been about 4,000,000. Thus, our CPB estimates reflect the long-term health benefit to recent birth cohorts, which in many cases is higher than the benefit to the current population.

Second, the birth cohort approach is the final piece in assuring that CPB captures the total benefit of the service. This is the case because, for screening services, the benefit is calculated for all individuals, not just those who are asymptomatic. The age for beginning the deliver of screening services as recommended by the USPSTF typically corresponds to the age at which onset of illness becomes more common. Therefore, very few individuals are symptomatic at the time screening begins. To estimate the benefit of the service to the current cross-section we would need to exclude those who are already symptomatic as a result of missed screening opportunities.

The birth cohort approach also defines the time frame for service delivery. The CPB and CE for each service is calculated for the age range recommended by the USPSTF or ACIP. Thus, the analysis for a screening intervention that is recommended from ages 40 to 70 would begin at age 40 and end when no additional benefits from the last screen age 70 are expected to be realized. Survival rates from US life tables are used to determine the portion of a birth cohort that is still alive at each screening age. Any prevented disease consequences that are not realized after the age of 70 from prior screening are included. Likewise, the benefits of counseling services are estimated for repeated counseling over the recommended age group.

In turn, this approach necessitates the choice of frequency of delivery. Not all services have an evidence-based recommended delivery interval. The USPSTF, for example, was unable to define an optimal interval for cervical cancer screening based upon the available data. In such cases, we reviewed guidelines of major professional associations and adapted the most commonly recommended interval.

In most cases, recommended delivery intervals are based upon information from effectiveness studies. Most effectiveness studies are conducted over 4 to 10 year time frames, providing adequate approximations of the effectiveness of the intervention when provided over the recommended age range of a birth cohort. Therefore, it is usually possible to identify effectiveness estimates that are consistent with our chosen delivery intervals and time frame. For a few services, the literature neither provides clear guidance on appropriate recommended intervals nor effectiveness estimates which reflect reasonable delivery intervals over the recommended age range. Brief tobacco cessation counseling, for example, is typically studied for one to three brief interventions over a time period that is rarely longer than one month. There is very little evidence on the effectiveness of brief tobacco cession counseling spaced months or years apart over an extended time frame. In such cases we derive an estimate from all available data that is consistent with recommended delivery intervals. The precision of such estimates is substantially less than effectiveness estimates that are observed or derived more directly, and is reflected in the sensitivity analysis for the service. The sources of all effectiveness estimates will be provided in each service's technical report.

4.6 Pairing Burden of Disease and Effectiveness Data

Conceptually, CPB is the burden addressed by the service multiplied by the effectiveness of the service. When combined in this way, effectiveness has a very specific definition: the percent reduction in burden addressed by the service. The definition of burden addressed by the service ('addressable burden'), on the other hand, is open to interpretation. For example, does cholesterol screening address all coronary heart disease (CHD), CHD caused by high cholesterol, CHD among individuals with high cholesterol (regardless of cause), or only the portion of CHD that can be successful prevented with medication or lifestyle modification? Does folic acid chemoprophylaxis address all neural tube defects or only those that appear to be caused by insufficient folic acid as determined by folic acid supplement studies?

In our experience, it not possible to choose a single definition of addressable burden that can be applied consistently across all services. However, the only requirements that must be met to maintain consistency in CPB across services are that addressable burden include the entire burden that can be prevented and a corresponding effectiveness estimate is available. For example, suppose that the baseline number of CHD deaths in a population is 1,000 and 40% of the CHD deaths are attributable to high cholesterol. If cholesterol screening is 50% effective in reducing the CHD deaths that are attributable to high cholesterol and 0% effective in reducing other CHD deaths, then it must be the case that cholesterol screening is 20% effective (40% x 50%) effective in reducing total CHD risk. It would not matter whether CPB is calculated as a 50% of the 400 CHD deaths attributable to high cholesterol or 20% of the 1,000 total CHD deaths. Either way, CPB for cholesterol screening would include 200 prevented CHD deaths.

In practice, there are very few services for which there is a choice of effectiveness estimates. As a result, the available effectiveness data typically dictate the measurement of addressable burden for each service. Because the scope of burden on which the effectiveness data are available differ from service to service, so does the scope of burden used in each services CPB calculation. As a result, direct comparisons of burden addressed by the service or percent effectiveness cannot be made even though comparisons of CPB are valid.

4.7 Measuring harms and side effects

We measure harms and side effects in the same manner as burden of disease: in QALYs, and we subtract the QALYs lost to harms and side effects to CPB to derive 'Net CPB'. However, due to poor data availability, we generally exclude from our calculations any harms that are so small in terms of frequency or severity to be inconsequential to the ranking. For example, localized reactions to immunizations at the injection site are greatly outweighed by the years of life gained. The USPSTF only recommends preventive services for which the benefits clearly outweigh the harms. Consequently, few of the services we evaluate have harms that are material to our ranking.

In most cases, the data needed to accurately quantify the health impact of harms are of poor quality or entirely lacking. Using readily available estimates and a set of assumptions for other data points, we use sensitivity analysis to determine whether there is a plausible set of estimates for which the impact of harms is likely to substantially change the base-case CPB estimate. These decisions are sometimes made in conjunction with other potential impacts of the service for which data are lacking in order to assess whether the net impact is likely to change the CPB estimate.

5 CONSISTENT ESTIMATION OF CE

The inconsistent methods within the CE literature are well recognized. The Panel on Cost Effectiveness in Health and Medicine (PCEHM) convened to produce guidelines that promote better methodology and improve consistency among published CE studies.[19] Although methodology has improved since the broad dissemination of the Panel's recommendations, some inconsistencies remain in new studies and we reviewed older CE studies with more consistency issues. Therefore, we needed a methodology to produce CE estimates that are comparable across services.

The methodology also needed to account for the fact that no viable CE studies exist for many services. Given the large number of services involved in the Prevention Priorities studies, we sought a methodology that was less resource intensive than building detailed Markov models for each service while maintaining comparability of CE estimates across services. We identified three types of CE estimates that could be used as available while maintaining consistency across services. We first reviewed published CE studies to determine if any estimates from them could be used after simple inflation adjustment. We did not find any studies that met this criteria for the 2006 ranking, and therefore all estimates were derived from one of the following two methods.

We reviewed published CE studies to determine if their results could be used after making adjustments to reflect the PCHEM reference case methods and, if necessary, the delivery of the service as recommended by the USPSTF. In addition to inflation adjustments, the most common adjustments were the addition of costs for patient time and travel for clinic visits, and reduction in service costs to account for non-adherence. In addition, it was sometimes necessary to calculate at weighted average of results to reflect the use alternative intervention technologies or delivery intervals. It was common to use reported model results for cost and health outcomes to calculate CE ratios rather than relying on CE ratios tabulated by authors. In most cases this allowed us to estimate average CE ratios (i.e. the CE of the service relative to no delivery of the preventive service) when only incremental CE ratios were reported by study authors.

For most services, we were unable to identify an existing study that would provide the basis for an estimate of CE that was comparable to CE ratios for other services in the Prevention Priorities project. In such cases we developed our own estimates, with our CPB estimates providing the basis. We added a discount factor for future health effects to our CPB estimates, and gathered additional data on utilization to calculate a cost-effectiveness ratio. Unlike most of the published CE estimates, these estimates are not based on full Markov models (time-transition models). However, in estimating the CPB-based estimates we adhered to the principals of the reference case methods of the PCHEM to ensure the comparability of estimates across services. In some cases, existing CE studies that could not be used as a base-case estimate provided a substantial amount of cost and other data for our CPB-based CE estimates.

5.1 Issues Common to CPB and CE

Sections 4.2 through 4.8 address conceptual and practical issues in estimating CPB that are also applicable to deriving comparable estimates of CE. We recount them here in brief, adding detail that is specific to estimating CE. Many of the methods that we applied to CPB are commonly used in the CE literature and therefore their extension to our CE estimates is quite natural.

5.1.1 Measuring Health Outcomes and Harms

QALYs are the recommended metric for health benefits in the reference case (ref). However, it is recognized that in some cases the small increase in precision from obtaining quality of life weights to adjust estimates of LYs saved may not be justified. In the case of cancer screening services, CE studies typically report results in terms of QALYs saved in sensitivity analysis using rough estimates for quality of life, it at all. In estimating CPB we assessed the potential impact of quality of life adjustment in conjunction with potential harms of screening and found that the net impact was small and that it was not clear whether the change would increase or decrease CPB. Therefore, we determined that using limited data to incorporate quality of life adjustments was not likely to increase the precision of the base-case CPB estimates. We made the same determination for CE and we followed the literature in expressing CE results for cancer screening services in terms of dollars per LY saved rather than dollars per QALY saved.

5.1.2 Defining Target Population and Timeframe

We used the same standard rule for defining the target population and timeframe in estimating CE as we used for CPB. We measured the CE of providing the service to a U.S. birth-cohort of 4,000,000 over the age range the service is recommended by the USPSTF at the recommended delivery interval.

The majority of published cost-effectiveness studies also use the birth-cohort approach. In screening studies, where the baseline prevalence of disease or risk-factor differ across age groups, the cross-sectional approach may provide a poor approximation of the CE of repeated screening in a birth cohort over time because the condition or risk factor that screening seeks to detect will be more prevalent at baseline. Therefore, we generally excluded CE studies based upon an analysis of a cross-section.

Many cost-effectiveness studies include a variety of delivery intervals in their analysis, making it possible to identify or calculate a CE ratio for an interval that closely matched the recommended interval. When no such CE ratio was available, we calculated our own CPB-based CE estimate. For example, estimates of the CE of long-term repeated tobacco cessation counseling are not available. Although estimates of one-time interventions are available for tobacco cessation counseling, it was necessary to build a CPB-based estimate of the CE of repeated counseling to maintain consistency across services in the Prevention Priorities project.

5.1.3 Accounting for Adherence

Incomplete adherence with offers to receive a preventive service has only a minor effect on CE ratios. Individuals who decline a service receive no health benefits and incur no service costs. The costs of clinician time to offer the service are lost, but these costs are small and are usually not included in CE estimates. Incomplete adherence at later stages, such as lack of adherence with follow-up for positive test results, can have a substantial impact on CE ratios because some resources have been invested and there can be no off-setting health or financial benefits when follow-up does not occur.

Occasionally, a CE study will fail to account for incomplete adherence. When high quality studies that account for incomplete adherence are available, we excluded other studies. When the only available CE study fails to account for incomplete adherence, we made case-by-case decisions on whether to make an adjustment to the published CE ratio (if possible), use the published CE ratio without adjustment, or create our own CPB-based CE estimate. In many cases, authors did not account for incomplete adherence because they have made a reasonable determination that it would not materially affect the results. Using such ratios without adjustment does not compromise the comparability of CE across studies. For the 2006 ranking, two CE studies were used with adjustments made to improve consistency in how non-adherence was treated in the costs and health benefits comprising the CE ratio.

5.1.4 Accounting for Current Delivery Rates

In measuring of CPB, we estimate the total benefit of the service for those currently receiving the service as well as those not yet receiving the service (see Section 4.4). The same principal was used in estimating CE: our estimates reflect the cost-effectiveness of offering each service to all individuals, not just those who are currently not receiving the service. Implicitly, this is the method of most CE studies of preventive services. They are typically based on an average risk population and thus representative of both those receiving the service and those not.

It also important to note that, unlike many literature estimates, our estimates reflect the 'average' CE ratio, rather than the incremental CE ratio of changing from one technology to another. CE studies that focus on incremental results also report average CE ratios or report results in enough detail to allow average CE ratios to be calculated. For this reason, the results we used are often not the main results reported in the article abstract and thus may seem to contradict published literature.

5.2 Inflation Adjustment

To improve the comparability of CE estimates across services, we inflation-adjusted all estimates to a common base-year dollar using the medical care component of the consumer price index (MCPI). To reduce the potential for introducing error from inflation adjustment we chose a recent, but not too recent year for our base year: 2000. It is well-documented that the MCPI is an imperfect measure of general medical price inflation and its accuracy for updating the price of any particular service, such as a vaccine or its administration, varies from service to service. Systematic errors in the inflation adjustment are magnified over the number of years for which inflation adjustment is made. By using year 2000 as our base, we introduce less error when adjusting older CE studies than if we had used a more recent year.

Some otherwise useful CE studies are limited due to the age of the cost data which underlie the estimates. Even some recent studies are based upon cost data that were originally measured in the early and mid 1980s when DRG payments for inpatient stays started to alter the utilization of hospital services. Simply updating these cost data for inflation could introduce significant error to the CE ratio. We excluded these studies from our estimates of the CE of the service if other high quality studies are available. When only studies using more older cost data were available we made a decision on a case-by-case basis as to whether a CPB-based CE model would produce a substantially better CE estimate than an inflation-adjusted estimate using the existing study.

5.3 Discount Rate

The PCHEM recommends using a 3% discount rate for the reference case to improve comparability of CE ratios. Most studies published since 1998 use a 3% discount rate, but older studies may use discount rates from 5% to 10%, and a few do not discount health benefits.

To improve comparability across studies we either select higher utility studies that used a discount rate or 3%, or, if necessary and possible, made adjustments to a published CE ratio that used a different discount rate. This was possible when the CE study reports sufficient detail to allow us to estimate the median year(s) into the future in which the costs and benefits occur. We developed present value tables that compare the effects of various scenarios. For example, by referring to the tables we were able to estimate the impact on the CE ratio of discounting deaths that occur a median ten years in the future at the average age of 85 using a 3% discount rate rather than a 5% discount rate. These tables also provided the basis for discounting in our CPB-based CE estimates. Although these estimates are not as precise as discounting in a time-transition model, our experience has shown that CE ratios are insensitive to wide variations in the estimates of median year that determine present value of future costs and benefits.

5.4 Societal Perspective and Productivity Losses

The PCEHM recommends using the societal perspective for the reference case to improve comparability across studies.[19] When using the societal perspective all costs and benefits, no matter who incurs the costs, are included in the analysis. However, the PCECHM also recommends excluding productivity gains and losses associated with diseases targeted by the service accept under special circumstances related to how QALYs are calculated.

Few recent studies include productivity losses in their CE estimates due to this recommendation and the difficulty of estimating productivity losses. Studies that do include productivity losses usually provide alternate estimates that exclude them to facilitate comparisons with others studies. We selected the estimates from studies that exclude productivity losses when they are reported.

The PCEHM's recommendations for the reference case should not be interpreted to mean that time lost and travel costs in receiving the services and treatments should be excluded. Unfortunately, few CE studies of preventive services include these costs. When studies provide sufficient detail, we added these costs into the numerator based upon average hourly compensation for the US population.[20] We used this rate for all persons, in the formal work force or not, as a proxy for the value of non-paid labor and leisure time. We assumed that an average office visit requires 2 hours of time including travel to and from the visit. For services that are likely to be provided as part of a visit during which other services are provided, such as influenza vaccinations for older adults, we attributed only a fraction of this time to the preventive service. The exact amount included is reported in each service's technical report.

It may seem contradictory to include time costs to receive services while excluding time costs for time saved through prevention. There is general agreement that including time costs for years of premature death preventing results in double counting of the value of years of life gained: once in the numerator of the CE ratio measured in dollars and once in the denominator of the CE ratio measured in years of life within the QALY metric. There is less agreement that including the value of time lost due to illness results in double counting. The PCEHM recommendation is rooted in the belief that, unless specifically instructed not to do so, individuals who participate in the utility rating exercises that yield QALY weights factor in time lost to illness in their responses. If this is the case, productive time lost to illness and disability is already reflected in the denominator of the CE ratio. There are no empirical data which indicate whether or not this is indeed the case. Although this PCEHM reference case recommendation has not been as widely accepted as others, we followed the recommendation because there is not a clear-cut rationale for making an exception to our principle of following PCHEM reference case methods.

The NCPP struggled with determining the best approach for including the time costs for behavioral interventions that require investments of patient time not spent with health professionals. While this may occur for all services, it is usually a minor cost and therefore not included in the estimates. For example, the value time spent to stop at the pharmacy to refill an cholesterol control medication is very small compared the costs of screening and the medication itself, and is virtually zero for those using mail-order systems for refills. In contrast, the USPSTF recommendation for obesity screening is based upon evidence that intensive interventions aimed at changing eating and physical activity habits following a positive screen are effective. Intensive interventions as investigated in the literature typically require a substantial amount of time over a one year period face-to-face with health professionals (dieticians, personal trainers, and other behavioral specialists) and the intended health benefits of screening will not be realized without continued time investment for physical activity.

The NCPP discussed several issues regarding the best approach to valuing of time spent in such activities. What activities did the those who adhere to recommended changes give-up in order to devote time to lifestyle modifications? What income and what enjoyment or other benefits did they derive from that activity? What enjoyment and other immediate benefits do those who adhere to recommendations derive from lifestyle modification (the long-term health benefits are captured in the QALYs saved in the denominator of the CE ratio)? What time-costs would patients want decision-makers to consider when choosing which preventive services to offer first? What approach would make the ranking most readily understood? What are the recommendations of expert panels and the literature precedents for valuing of time costs of life-style modification?

A complete discussion of these topics and their complex inter-connections is beyond the scope of this document, but selected comments may be particularly instructive. We found the CE literature on life-style modifications to be sparse and did not provide a clear example to follow. Some studies included no time costs for lifestyle modifications(ref) while others included some costs while excluding others without providing a rationale that is well-founded in economic theory.(ref) The general PCEHM reference case recommendations are clear in stating that all time costs should be included in all analyses from the societal perspective, but their very brief discussion of implementing time costs in practice for physical activity is inconclusive.[19]

In a majority vote, the NCPP chose to include time costs associated with receiving services (face-to-face interaction with health professionals), and to exclude other time costs, such as those for long-term physical activity maintenance that is not performed in the presence of a personal trainer. The reasons for choosing this approach included consistency with other services in the ranking for which the only time costs included involved face-face interaction with health professionals and travel time to such visits; the belief that decision-makers should leave the valuation of other time costs to the patient; and the possibility that the small minority of patients who adhere with long-term life-style modifications do so because they receive short-term benefits that are at least equally valuable to them as the activities they gave up.

5.5 Using Evidence from Multiple CE Studies

Very few services have several high-utility CE studies available in the literature. When there was more than one study, it was usually possible to identify an article that better suits our needs due to better accounting for one or more quality issues, better alignment with the service technology or target population with that recommended by the USPSTF, or more comprehensive reporting that allowed us to recalculate results to improve consistency of CE estimates across services by making adjustments such as those noted above and to perform more complete secondary sensitivity analysis.

Therefore, we did not average results of multiple CE studies to derive our CE estimates. Instead, we identified the best study based primarily upon its concordance with the preventive service and target population as recommended by the USPSTF, the agreement between the studies methods and the PCECHM reference case methods and the detail of reporting. In sensitivity analysis for the CE of colorectal cancer screening, we recalculated an adjusted CE ratio based upon a second CE study that we judged to be nearly equal in applicability to the study that was chosen for our base-case estimate. The difference between the adjusted results was too small to affect the ranking.

5. RANKING SERVICES

The base-case CPB and CE estimates are subject to the uncertainty of the underlying data, and to a lesser extent, some simplifying calculations. Even services with the best data, care must be taken in basing decisions on a single point estimate. A ranking with one service receiving higher priority than another based on small differences in base-case estimates for either CPB or CE would be built on false precision. The base-case estimates are still informative because there is a wide range of base-case estimates across all services for both CPB and CE, even though the base-case estimates are close for some services. The goal of the ranking is to aid decision-making by defining broad groupings of services with similar value among these wide ranges.

A scoring system was used to achieve this goal without overstating the precision of the CPB and CE estimates. First, services were sorted in descending order by the CPB base-case estimates and in ascending order by the base-case CE ratios. Each service was then assigned a score from 5 to 1 for both CPB and CE, according to groups split at the quintiles of range of base-case estimates. Services with the highest CPB were thus assigned a CPB score of 5, and services with the lowest CE ratios (or with cost-savings) were assigned a CE score of 5. Scores for CPB and CE were then added to give each service a total possible score between 2 and 10.

This scoring system embodies several judgments. First, the NCPP chose quintiles for scoring categories to balance the goal of providing information about differences in the services relative value against the risk of creating an inaccurate ranking due to imprecise base-case estimates. Broader categories such as those created by quartiles or tertiles would reduce the risk of an inaccurate ranking, but would hide useful information about differences in the base-case estimates. Finer categories, of course, would reveal more about the differences in base-case estimates, but would introduce a larger risk for inaccurate rankings.

Our sensitivity analysis revealed that many scores might be different from the assigned score by one when using quintiles (i.e. assigned a score of 4 for CPB, but the ‘true’ CPB could be consistent with a score of 3 or 5), but few scores would differ by more than one. The situation would have been similar when using quartiles. The importance of the potential for difference in score of one should be interpreted with the sensitivity analysis methods in mind (see section 6). The potential for scoring errors were assessed using the ranges defined in multivariate sensitivity analysis. For the ranges defined by sensitivity analysis to better reflect the true value of the service than our base-case estimate, we would have had to simultaneously misstated the three most influential variables a significant amount in a direction that is either more or less favorable to the CPB or CE estimate. In our view, an error in score of one is possible for nearly all scores, but is only likely to occur when base-case estimates are close to the quintiles, important data points are particularly uncertain, or the model is particularly unstable due to the nature of the preventive service, with the later scenario occurring only with CE scores.

Another important point in the choice of groupings is the decision to use cut points defined by quintiles rather than defining cut points by natural breaks in the distributions or some other qualitative criteria. The use of natural breaks could have reduced the potential for base-case estimates to fall near the cut-off points and thus reduce the potential for scoring errors. However, the objectivity and transparency of quintiles compensates for this shortcoming.

The simple addition of CPB and CE scores to derive a total score results in both criteria being given equal weight. There is no theoretical foundation for combining these criteria into a final score, or for giving them equal weight when doing so. Therefore the simple and transparent approach of adding scores with equal weight was chosen to facilitate interpretation of the ranking. The sorting of CE and CPB created ordinal scales. Addition of ordinal scales may result in spurious total scores. For example, it is possible that investments to improve the delivery rates of a service with a CPB-CE score combination of 3-4 (total = 7) is a less desirable than investments in a service with a score combination of 4-2 (total = 6). This is because the move along the CE scale associated with a decrease of the CE score from 4 to 2 may be less important than a move along the CPB scale that increases that score by one (from 3 to 4). For the total score to accurately reflect best practices, it would have to be the case that CPB and CE are equally valued in decision-making. There is no reason to believe this is true for every decision-maker. Therefore, while the total ranking provides a convenient starting place, each decision-maker needs to pay attention to the underlying scores and their uncertainty.

6. SENSITIVITY ANALYSIS

We conducted sensitivity analysis for CPB and CE for each service. The primary goal of the sensitivity analysis was to determine the possible range of scores for each measure. Unless noted below, all methods described here apply to both CPB and CE, but sensitivity analysis was conducted separately for each. Thus, the variables that were determined to be most influential for CPB are not necessarily among the most influential variables for CE.

For all variables we defined plausible ranges for each data point using the base case as the midpoint. The sensitivity analysis was designed to determine the possible range of estimates for CPB and CE for a US birth cohort. Therefore, the ranges reflect the plausible range of national average, given uncertainty in the available data. They do not represent variation that might be observed among health plans, clinics, or individual patients. Variation at these levels may be substantially higher than the plausible values for the national average.

We conducted both single-variable and multiple-variable sensitivity analyses to identify the variables that collectively cause the largest change in CPB and CE estimates. We first used single-variable sensitivity analysis to determine the individual variables that cause the largest changes. Single variable sensitivity analysis typically identified 4 to 7 variables to which the estimates were most sensitive. We tested all combinations of these variables to identify the three variables that produced the largest change in each base-case estimate when the variables were simultaneously changed in a less favorable direction. This step was repeated changing the variables in the more favorable direction. The resulting less favorable and more favorable CPB and CE estimates were used to assess the robustness of the scores.

Several considerations went into conducting this sensitivity analysis. First, CPB and CE estimates are derived from variables that may be related to one another. For example, a service may address multiple causes of mortality, such as tobacco cessation counseling, and the mortality data for each cause is likely to be collected in the same manner. In cases such as this, an under- or over-estimate of the value of one variable due to imperfect source data may also effect other related variables and cause a systematic bias. While measurement error of any one of the variables may not substantially influence the estimate of CPB or CE, measurement error in all related variables may.

Therefore, to guard against systematic error in estimates we grouped related variables (e.g., total discounted QALYs lost, total discounted intervention costs, total discounted costs of illness, and total adherence) and treated the group as an individual variable. If a group variable was more influential than any one of the three "most influential" variables from the first series of sensitivity analyses, it was substituted for the least influential of the most influential variables. For CPB, we analyzed the potential for systematic error in the components of QALYs lost (mortality incidence, average life expectancy, morbidity incidence, morbidity duration, and morbidity QALY weight), rather than total QALYs lost.

CE estimates based on adjustments to published estimates required a different strategy for sensitivity analysis because we did not reconstruct the model from each data point. Also, it was not possible to use each study's reported sensitivity analysis because they did not report sensitivity analysis for all data points and they frequently reported sensitivity analysis in terms of incremental CE ratios, not the average CE ratios used in the ranking. It was not possible to define a strict set of procedures for sensitivity analysis for adjusted CE ratios because the reporting of results and number of adjustments varied by service. The following general rules were followed, but decisions with respect to the exact method followed for each circumstance were guided by the goal of creating an uncertainty ranges for CE estimates that were consistent across services. For adjusted CE ratios from published studies, we estimated ranges based on the two most influential variables among 'aggregate' variables that are often reported in results: total discounted costs, total discounted savings, and total discounted QALYs saved. The sensitivity of the CE estimates to our adjustments were tested along with these summary variables. When two or more adjustments were made, or when the CE ratio was found to be particularly sensitive to one adjustment, then the range for sensitivity analysis was based upon two aggregate variables plus variation in the most influential adjustment parameter.

Reference List

1. Coffield AB, Maciosek MV, McGinnis JM, Harris JR, Caldwell MB, Teutsch SM, Atkins D, Richland JH, Haddix A. Priorities among recommended clinical preventive services. Am J Prev Med 2001 Jul;21(1):1-9.

2. U.S. Preventive Services Task Force. Guide to clinical preventive services, 2nd ed. Baltimore: Williams & Wilkins; 1996.

3. Maciosek MV, Coffield AB, McGinnis JM, Harris JR, Caldwell MB, Teutsch SM, Atkins D, Richland JH, Haddix A. Methods for priority setting among clinical preventive services. Am J Prev Med 2001 Jul;21(1):10-9.

4. Holtgrave DR. Extending the methodology of the Committee on Clinical Preventive Service Priorities to HIV-prevention community planning. Am J Prev Med 2002 Apr;22(3):209-10.

5. Vogt TM, Aickin M, Ahmed F, Schmidt M. The Prevention Index: using technology to improve quality assessment. Health Serv Res 2004 Jun;39(3):511-30.

6. Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright-De Aguero L, Truman BI, Hopkins DP, Mullen PD, Thompson RS, Woolf SH, et al. Developing an evidence-based Guide to Community Preventive Services--methods. The Task Force on Community Preventive Services. Am J Prev Med 2000 Jan;18(1 Suppl):35-43.

7. Carande-Kulis VG, Maciosek MV, Briss PA, Teutsch SM, Zaza S, Truman BI, Messonnier ML, Pappaioanou M, Harris JR, Fielding J. Methods for systematic reviews of economic evaluations for the Guide to Community Preventive Services. Task Force on Community Preventive Services. Am J Prev Med 2000 Jan;18(1 Suppl):75-91.

8. Zaza S, Wright-De Aguero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, Sosin DM, Anderson L, Carande-Kulis VG, Teutsch SM, et al. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services. Am J Prev Med 2000 Jan;18(1 Suppl):44-74.

9. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1993 Dec 1;270(21): 2598-601.

10. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. Jt Comm J Qual Improv 1999 Sep;25(9):470-9.

11. Slack MK, Draugalis JR. Establishing the internal and external validity of experimental studies. Am J Health Syst Pharm 2001 Nov 15;58(22):2173-81; quiz 2182-3.

12. Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R, Lam M, Seguin R. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003 Dec 22;3:28.

13. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii-x, 1-173.

14. Screening for colorectal cancer: recommendation and rationale. Ann Intern Med

2002 Jul 16;137(2):129-31.

15. Screening adults for lipid disorders: recommendations and rationale. Am J Prev Med 2001 Apr;20(3 Suppl):73-6.

16. Pignone MP, Phillips CJ, Atkins D, Teutsch SM, Mulrow CD, Lohr KN. Screening and treating adults for lipid disorders. Am J Prev Med 2001 Apr;20(3 Suppl):77-89.

17. Murray, C. J. L.; Lopez, A. D. The Global Burden of Disease: Volume I. A comphrehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. United States of America: The Harvard School of Public Health on behalf of The World Health Organization and The World Bank; 1996. (Global Burden of Disease and Injury Series.

18. Murray, C. J. L.; Lopez, A. D. Global health statistics: Volume II. A compendium of incidence, prevalence and mortality estimates for over 200 conditions. United States of America: The Harvard School of Public Health on behalf of The World Health Organization and The World Bank; 1996. (Global Burden of Disease and Injury Series.

19. Gold, M. R.; Siegel J. E. ; Rusell L. B. , et al. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996.

20. Employer Costs for Employee Compensation Historical Listing (Annual), 1986-2001. 2002 Jun 19.

Figure 1A. Standardized Search Levels

Topics: Effectiveness/Cost Effectiveness

Revised 01/30/04

Level 1	Level 2	Level 3*	Level 4*
Search PubMed Limit to English language. Limit to MeSH major terms, title word terms, and phrases. Back to 1992 (01/01/92). Exclude publication types – editorial, comment, news and letter. Search Cochrane back to 1992. Obtain systematic review articles published back to 1992. Obtain articles used as part of the review that were published back to 1987. References from major articles identified in Level 1.	Search PubMed Limit to English language. Limit to text word terms. Back to 1992 (01/01/92). Exclude publication types – editorial, comment, news and letters. References from major articles identified in Level 2.	Search PubMed Limit to English language. Limit to MeSH major terms, title word terms, and text word terms. Back to 1987 (01/01/87). Exclude publication types – editorial, comment, news and letter. Other knowledge-based informational databases (literature databases).	Search PubMed Limit to English language. Limit to text word terms. Search general web. Association websites (American Heart Association, American Cancer Society, etc). Search PubMed for English abstracts from all languages.

*Search one or more of the options listed in Level

Figure 1B. Standardized Search Levels
Topics: Burden of Disease and Cost

Revised 01/30/04

Level 1	Level 2	Level 3	Level 4*
National data sets Government websites (CDC, NIH, AHRQ, etc.) Search PubMed Limit to English language Limit to MeSH major terms and title word terms. Back to 1998 (01/01/98). Exclude publication types - editorial, comment, news and letter. Data sources referenced by articles identified in Level 1.	Search PubMed Limit to English languagea Limit to MeSH terms and phrases. Back to 1998 (01/01/98). Exclude publication types – editorial, comment, news and letter Association websites (American Heart Association, American Cancer Society, etc.) Data sources referenced by articles identified in Level 2	Search PubMed Limit to English language. Limit to MeSH major terms, title word terms, MeSH terms and phrases. Back to 1990 (01/01/90). Exclude publication types – editorial, comment, news and letter. Data sources referenced by articles identified in Level 2.	Search PubMed Limit to English language. Limit to text word terms (Search only as appropriate i.e., if it would represent current health status). Other knowledge-based informational databases (literature databases). Search general web HealthPartners' Data

* Search one or more of the options listed in level.

AHRQ, Agency for Healthcare Research and Quality; CDC, Centers for Disease Control and Prevention; NIH, National Insititutes of Health.

Figure 2.
Utility Score Categories

Effectiveness	Cost effectiveness
Study Population Consistency with recommended service age group Generalizability to U.S. population	**Study Population** Consistency with recommended service age group Generalizability to U.S. population
Service Definition Consistency with USPSTF technology Consistency with recommended frequency of delivery	**Service Definition** Consistency with USPSTF technology Consistency with recommended frequency of delivery
Reporting Completeness and clarity	**Reporting** Completeness and clarity
Design Adequacy of study design and study measures to generate reliable estimates Adequate sample size	**Design** Consistency with PCEHM reference case methods
Implementation Consistent with design	**Sensitivity Analysis** Completeness and clarity of sensitivity analysis in identifying variables influential to average CE ratio.
Evaluation/Analysis Appropriate statistical analysis Adequate control of potential selection bias and other population differences	

CE, cost effectiveness; PCEHM, Panel on Cost Effectiveness in Health and Medicine; USPSTF, U.S. Preventive Services Task Force.