NATIONAL QUALITY FORUM

TO: NQF Members

FR: Karen Pace, PhD, RN

RE: Review of *Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement*

DA: May 20, 2010

Background

NQF's evaluation criteria require a variety of evidence including the clinical evidence for the focus of a quality measure (criterion 1c). Steering committees have diverse backgrounds and expertise and could benefit from more guidance and support to consistently apply the NQF measure evaluation criteria. Both evidence and expert judgment play a role in evaluating measures against criteria, however, judgment can best be applied when Steering Committees have a thorough understanding of the evidence that does or does not exist.

In January, NQF convened a task force of seven members to assist with developing guidance on evaluating the evidence that supports the focus of a quality performance measure (1c), as well as the other subcriteria under *Importance to Measure and Report*. The task force was asked to address the following tasks.

- Identify the type of evidence needed to justify the focus of a quality measure (1c) (i.e., what is being measured).
- Identify the evidence needed to demonstrate high impact (1a) and opportunity for improvement (1b).
- Develop guidance on how technical advisors and steering committees use the evidence provided to evaluate submitted measures for possible endorsement.
- Make recommendations for potential enhancements to the evaluation criteria.

The Task Force's recommendations are included in the draft document, *Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement*. The draft report is posted on the NQF web site for review and comment only—not voting.

You may post your comments and view the comments of others on the NQF website. NQF is now using a program that facilitates electronic submission of comments on this draft report. **All comments must be submitted using the online submission process.**

NQF Member comments must be submitted no later than 6:00 PM ET, June 18, 2010; public comments are due 6:00 PM ET, June 11, 2010.

Supporting documents related to your comments may be submitted by e-mail to performancemeasures@qualityforum.org with "Evidence Report" in the subject line and your contact information in the body of the e-mail.

Thank you for your interest in the NQF's work. We look forward to your review and comments.

NATIONAL QUALITY FORUM

Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement

Draft Report for Review and Comment

May 20, 2010

NATIONAL QUALITY FORUM

Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement

Draft Report May 20, 2010

CONTENTS

OVERVIEW AND PURPOSE

Steering committees have diverse backgrounds and expertise and could benefit from more guidance and support to consistently apply NQF measure evaluation criteria. Both evidence and expert judgment play a role in evaluating measures against criteria. However, judgment can best be applied when Steering Committees have a thorough understanding of the evidence that does or does not exist. Evidence comes in many different forms (e.g., peer reviewed publications; practice guidelines from authoritative sources; expert assessments); there are often inconsistencies and gaps; and it can be difficult to interpret and reach conclusions. In October2009, the Board directed that NQF should take steps to strengthen its processes to evaluate the synthesis and scoring of evidence and to present this information in ways that will be best understood and useful to Steering Committees.

NQF's evaluation criteria require a variety of evidence as noted in the following table. Of these criteria, some of the most rigorous evidence is required to justify what is being measured (1c) and that is the primary focus of this report – ***the evidence required to justify the measure focus*** (i.e., the specific process, structure, outcome, etc. that is being measured). Another task force and subsequent report will address measure testing and the criterion of *Scientific Acceptability of Measure Properties*.

NQF endorses measures that are intended for use in public reporting as well as quality improvement with the goal of improving the quality of healthcare. The evidence that supports the focus for a quality measure is addressed under the must-pass criterion, *Importance to Measure and Report* because if the measure focus is not supported by evidence that it can facilitate gains in quality and health, then the use of limited resources for measuring and reporting on it would be questionable. For most healthcare quality measures, the evidence will be that of clinical effectiveness and the link to desired outcomes.

Table 1. Measure Evaluation Criteria and Type of Evidence

Evaluation Criteria	Type of Evidence
1. Importance to measure and report 1a. High impact	Epidemiologic data
1b. Opportunity for improvement	Epidemiologic data Health services research
1c. Evidence that supports the focus of measurement	Clinical research; Health services research
2. Scientific acceptability of measure properties 2a-h	Psychometric testing - reliability and validity, adequacy of risk adjustment, etc.
3. Usability 3a. Demonstration of understanding and usefulness for public reporting and quality improvement	Data and/or qualitative information demonstrating usefulness for public reporting and quality improvement Understanding what? (ECS)
4. Feasibility 4e. Demonstration the measure can be implemented	Data and/or qualitative information demonstrating the measure can be implemented

Task Force Charge

The task force was asked to address the following tasks.

- Identify the type of evidence needed to justify the focus of a quality measure (1c) (i.e., what is being measured).
- Identify the evidence needed to demonstrate high impact (1a) and opportunity for improvement (1b).
- Develop guidance on how technical advisors and steering committees use the evidence provided to evaluate submitted measures for possible endorsement.
- Make recommendations for potential enhancements to the evaluation criteria.

BACKGROUND

Ideally, quality performance measures are based on high quality evidence regarding the types of interventions and services that will achieve desired outcomes and reflect high quality care. However, much of healthcare has not been subjected to research studies, much less with randomized controlled trials or comparative effectiveness studies. Lohr observed that "Perhaps

no more than half, or even one-third, of services are supported by compelling evidence that benefits outweigh harms [1]." Many quality performance measures are based on clinical practice guidelines, however not all guideline recommendations are appropriate for performance measure development, which depends on the strength of the evidence and relationship to meaningful outcomes [2]. For example, Tricoci, et al. [3] reviewed recommendations in American College of Cardiology/American Heart Association guidelines and found that only 314 of 2711 recommendations were classified as A-level evidence based on multiple randomized trials with large numbers of patients.

Some aspects of healthcare (e.g., system change) may be more difficult to study with quantitative methods, particularly with randomized controlled trials. Some clinical process steps (i.e., assessing health status, diagnosing clinical conditions, recommending treatment, teaching and counseling about conditions/treatment) may be unlikely to be subjected to research. Even when research has been conducted, the body of evidence may not have been systematically assessed and graded (e.g., care coordination, medication management). Lohr [1] noted that absence of evidence about benefit is not the same as evidence of no benefit. Even when available, evidence is rarely definitive. However, the level of confidence in a recommendation (or measure) depends on the underlying research and synthesis of that research.

Evidence Issues Identified with Measures Submitted to NQF

The NQF evaluation criteria (1c, Footnotes 3 & 4) and submission questions may not provide enough direction to reviewers or measure developers. Measure submissions often have insufficient information on the strength of the evidence or strength of a guideline recommendation. Measures have been submitted with no evidence; no systematic grading or incorrect grading of the evidence or guideline recommendation; use of a different grading system than the recommended USPSTF system with no explanation; or low quality evidence. In some cases, a grade might be assigned without using the associated methods to assess the body of evidence. Some submitted measures are focused on process steps far removed from the desired outcome, even when there is evidence for a particular intervention or intermediate outcome that is more directly linked to the desired outcome (e.g., measures to assess

immunization status rather than measures of administering the vaccine). Some measure submitters question whether the suggested USPSTF evidence grading system is only applicable to preventive services.

NQF consensus projects were not intended to undertake systematic evidence reviews for the variety of measures that are submitted for consideration, nor is this feasible. The responsibility for such reviews lies with measure developers. However, in the past, measure developers have varied substantially in the expertise and resources they have to systematically assess the strength of a body of evidence or crosswalk a different grading system to the USPSTF system. Such detailed evidence reviews have frequently not been viewed by developers as an integral part of the measure development process. NQF wishes to clearly signal, through this document and the measure submission form itself, that evidence reviews need to be completed by measure developers or their designates prior to measure submission for endorsement.

The Changing Environment

As guidelines and quality metrics are increasingly used not only for internal quality improvement but also for public reporting, the necessity for a strong evidence base has become more urgent and compelling. This need is further substantiated by the development of reimbursement programs that utilize such publicly reported metrics. Although public reporting and pay for performance have the potential to inform consumers, focus quality improvement activities, and reward high performance; there are potential unintended negative consequences if measures do not meet all the aspects of the importance criterion. Potential negative consequences include confusion about the importance of particular care processes to quality, the unnecessary resources to measure elements of care that may not impact quality, and diversion of scarce resources to marginally effective activities. To achieve the intended positive effects of quality measurement and minimize the unintended potential negative consequences, measures should be based on the best evidence for the focus of measurement and also should conform to the highest measurement science principles. Recognizing the high stakes of performance measurement in an increasingly transparent environment, some measure developers have enhanced their requirements for the evidence base for performance measure development [4].

Clinical Practice Guidelines

Although they are not the only evidence base for performance measures, many measure developers rely on clinical practice guidelines to support the focus of measurement [2, 4]. There has been a proliferation of such guidelines, some overlapping or even contradictory. There also is substantial variability in the methodological rigor of review and grading of the evidence and recommendations. In 2000, Grilli [5] and colleagues reported that of 431 specialty society guidelines reviewed, 82% did not apply explicit criteria to grade the scientific evidence used as a basis for recommendations, 87% did not report whether a systematic literature search was conducted, and 67% did not describe the professional involved. Some tools to assess clinical practice guidelines [6-8] are available and developing trustworthy guidelines is also the subject of a current IOM study.

At the January 11, 2010 IOM meeting on developing trustworthy guidelines, Vivian Coates presented the following information about the National Guidelines Clearinghouse (NGC):

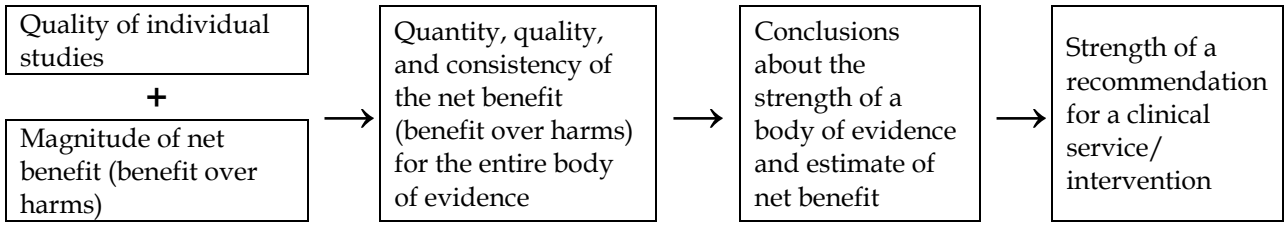
- Currently, NGC contains more than 2500 guidelines from more than 200 developers.
- Most of the developers whose guidelines are represented in NGC (158 of 204; 77%) use some sort of rating scheme to grade the underlying evidence and/or strength of the recommendations. Of these:
    - Ten developers report using GRADE or modified GRADE.
    - Six report using the USPSTF approach, either as is, or modified.
    - The great majority (142 developers) does not identify the origin of their rating schemes, and appear to be using schemes unique to their organizations.

Evidence Grading Systems

A variety of evidence grading systems currently are in use to achieve this enhanced degree of evidence review and assessment. These systems generally include methods for selection and review of the evidence, and rules or hierarchies related to grading the quality of evidence and the strength of a recommendation. These evidence grading systems are applicable to guidelines as well as other sources of evidence for performance measures.

There are commonalities among the various evidence grading systems. In general, the quality and strength of the overall body of evidence is a function of the *quantity* and *quality* of individual studies and the *consistency* among studies regarding judgments of net benefit (the balance of benefits and harms). *Quality* of individual studies includes study design, sample size and statistical power considerations, flaws such as selection bias, and generalizability of findings. Of particular interest for quality measures is how well the measure matches the population and intervention in the evidence (e.g., cited studies). The general approach to determining the strength of evidence and a recommendation for a particular intervention or service is depicted in Figure 1.

Figure 1. Approach to Determining Quality of Evidence and Strength of Recommendation



Differences in terminology and grading scales may inhibit understanding about the strength of evidence. Differences can range from a rather minor but understandable difference in terminology (e.g., strength, quality, or level of evidence) to pronounced differences in the assignment of grades (e.g., a grade of A could indicate evidence based on consensus of opinion in one system to evidence based on meta-analyses of randomized controlled trials in another system). An international initiative to standardize grading evidence and recommendations, **GRADE** [9-15], is now supported by many organizations including the Cochrane Collaboration. The Agency for Healthcare Research and Quality (AHRQ) supports two evidence grading systems: one used by the US Preventive Services Task Force (USPSTF) [16, 17] and one used by the Evidence-Based Practice Centers [18] (consistent with GRADE). Table 2 provides examples of terminology used by four evidence grading systems. It is important to note that grading systems are tied to specific methods for reviewing and assessing the quality of evidence.

Table 2. Examples of Terminology in Selected Grading Scales

	USPSTF	GRADE	AHRQ Evidence-Based Practice Centers	ACC/AHA
Evidence	Certainty of Net Benefit: • High • Moderate • Low Magnitude of Net Benefit: • Substantial • Moderate • Small • Zero/Negative	Quality of Evidence: (confidence in estimate of effect to support recommendation) • High • Moderate • Low • Very Low	Strength of Evidence: (confidence that estimate of effect is correct) • High • Moderate • Low • Insufficient	Estimate of certainty of treatment effect • A: multiple pop, RCT, meta-analysis • B: limited pop, single RCT or non-RCT • C: very limited pop, consensus expert opinion, case studies Size of treatment effect • Class 1: Benefit >>>Risk • Class IIa: Benefit >>Risk • Class IIb: Benefit > or = Risk • Class III: Risk > or = Benefit
Recommendation	Grade of Recommendation: Certainty/Magnitude • **A - Recommend**: High/Substantial • **B - Recommend**: High/Moderate; Moderate/Substantial; Moderate/Moderate • **C – Recommend against routine use**: High or Mod/Small • **D – Recommend against**: • High or Mod/Zero-Neg • **I-Insufficient evidence**: Low/any magnitude	Strength of Recommendation: • Strong • Weak	Does not make recommendation	• Should be performed: Class 1-A, B, C • Reasonable to perform: Class IIa-A,B,C • May be considered: Class IIb-A,B,C • Not helpful/may be harmful: Class III-A,B,C

Systematic reviews and meta-analyses are used to assess a body of evidence. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) focuses on the transparent and full reporting of such reviews [19]. The Institute of Medicine (IOM) has two consensus projects underway that relate to grading the quality of evidence for clinical interventions: Standards for Developing Trustworthy Clinical Practice Guidelines and

Standards for Systematic Reviews of Clinical Effectiveness Research; however, reports will not be ready until early 2011.

RECOMMENDATIONS

The Task force identified some principles that guided its discussion and the recommendations that follow.

Principles

Transparency is a primary goal. All stakeholders need to have a clear understanding of the evidence supporting a performance measure in order to make informed decisions about the importance of measuring and reporting on the topic.

Measures that will be used for public reporting should meet a high standard of evidence for the focus of measurement. NQF measures are intended to be useful for public reporting, as well as to internal quality improvement activities. Measures used for public reporting often impact large numbers of providers and entail investment of significant resources in measurement and improvement. Consequently, measures that will be used for public reporting should meet a high standard of evidence for the focus of measurement. A lower standard of evidence may be deemed appropriate by those selecting measures for use in smaller scale internal quality improvement activities within a learning system that allows for rapid adjustments.

In the absence of strong evidence of certainty of net benefit for the structure or process being measured, expert judgment must conclude that potential benefits to patients clearly outweigh potential harms to patients from the specific structure, intervention or service. Much of healthcare has not been subjected to research studies and thus, does not have a strong evidence base. In the absence of strong evidence, clinical interventions and services that are the focus of quality performance measures should be judged to have benefits to patients that clearly outweigh any potential risk. In the absence of strong evidence, administrative, management, or system structures and processes that are the focus of quality performance measures should be

judged to have benefits to patients that clearly outweigh the system costs and resources to implement those structures and processes.

Standards for evidence grading are evolving and expectations for both the present and future should be stated. Standards for evidence review and grading and clinical practice guideline development are evolving, as are expectations for measures endorsed by NQF. Explicit information about the evidence supporting a measure and how (or if) it was graded is essential for evaluating the evidence both now and in the future.

Consistency with prior terminology, whenever possible, minimizes confusion. Terminology used in prior NQF documents should be changed only if incorrect or leads to increased understanding. Whenever possible, narrative descriptions should be used instead of technical terminology.

I. Recommendations on Sources of Evidence and Evidence Grading for the Present and the Future

- The preferred sources of evidence are systematic reviews and grading of evidence conducted by independent organizations such as USPSTF, AHRQ Evidence-based Practice Centers, and the Cochrane Collaborative; or guidelines that meet national standards for trustworthy guidelines.
- Until such time when guidelines are certified to meet a set standard, preferred guidelines are those developed with balanced representation beyond one specialty group and with full disclosure of biases.
- An assigned evidence grade alone is not sufficient to evaluate whether the NQF criterion on evidence for the focus of measurement (1c) is met, either now or in the future. The specific information on the quantity, quality, and consistency of the body of evidence that was used to determine an overall grade should be provided in the measure submission.
- Explicit, transparent information on the quantity, quality, and consistency of the body of evidence supporting a measure will facilitate identification of guideline recommendations that do not have acceptable evidence as the basis for performance measurement. Explicit information about the evidence also facilitates review by all stakeholders although TAPs

and Steering Committees will continue to include experts that possess knowledge about the state of science for a particular topic.

- **Current Expectations –**
  - NQF should require measure developers to provide specific information about the quantity, quality, and consistency of evidence. Information should include how the evidence was graded and the grade assigned. If the developer fails to provide this information, NQF should not review the proposed measure.
  - NQF prefers that evidence be graded based on the systems of either the USPSTF or GRADE.
- **Future Expectations –**
  - Rather than identifying "preferred" grading systems as noted for the current expectations, NQF should require the use of one or two standardized evidence grading systems (e.g., the USPSTF, GRADE, or possibly one adopted by the IOM).
  - The evidence should be graded by identified credible sources, such as guideline developers or review organizations certified as meeting accepted standards.
  - Even with standardized grading systems and potentially certified reviewers, explicit information on the quantity, quality, and consistency of the specific evidence that led to the assignment of a grade should be submitted for evaluation. In other words, NQF expects not simply the end-result of the grading process, but also a concise summary of the evidence.

II. Recommendations for the Evidence Needed to Justify the Focus of a Quality Measure

There has been widespread acceptance of Donabedian's [20, 21] structure-process-outcome model for assessing healthcare quality. These three approaches to quality measurement can be used with any topic of healthcare quality and the evidence required generally does not vary by topic. The required evidence is for the links depicted by the red arrows in Figure 2. As depicted under process, there may be multiple process steps prior to delivering an intervention; however, the evidence is most often about the relationship between the intervention and outcome.

Figure 2. Structure-Process-Outcome Model

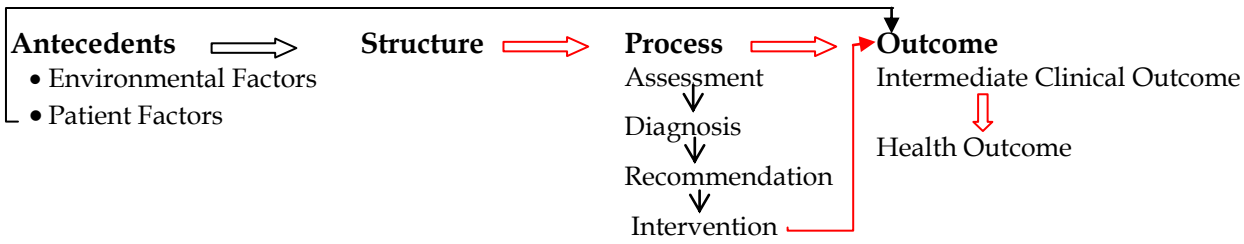


Table 3 outlines the evidence required to justify the structure, process, or outcome that is the focus of measurement (i.e., what is being measured). It also identifies special considerations related to certain quality topics. Subsequent tables lay out the approach for evaluating the evidence and using it to determine if the NQF criterion is met.

Outcomes as a representation of quality also are based on the process-outcome link. Outcomes are viewed as useful quality indicators because they are integrative of the influence of multiple care processes and disciplines involved in the care. However, that also presents some challenges related to presenting evidence to support the focus of measurement. Optimally, there will be a body of evidence for the link between the outcome and at least one care process. However, the lack of such evidence should not necessarily be reason to automatically dismiss the value of measurement, particularly when the outcome represents the reason for seeking and providing healthcare (e.g., health, function, survival, symptom control) or harm resulting from healthcare provided or omitted. Once outcomes are measured and reported, many outcomes that were not thought to be modifiable tend to be improved and stimulate identification and adoption of effective practices.

339 Table 3. Evidence to Support the Focus of Measurement

340

Type of Measure	Evidence	Example of Measure Type & Evidence to be Addressed
Structure Structure of care is a feature of a health care organization or clinician related to its capacity to provide high quality health care	Quantity, quality, and consistency of a body of evidence that the measured healthcare structure leads to desired health outcomes(including evidence for the link to effective care processes and the link from the care processes to desired health outcomes)	**#0190** Nurse Staffing Hours **Evidence** that higher nursing hours are associated with lower mortality, morbidity ; or associated with effective care processes (e.g., lower medication errors) that lead to better outcomes
Process A process of care is a health care-related activity performed for, on behalf of, or by a patient	Quantity, quality, and consistency of a body of evidence that the measured healthcare process leads to desired health outcomes in the target population If the measure focus is on inappropriate use: Quantity, quality, and consistency for a body of evidence that the measured healthcare process does <u>not</u> lead to desired health outcomes in the target population	**#0551** ACE Inhibitor / Angiotensin Receptor Blocker(ARB) Use and Persistence Among Members with Coronary Artery Disease at High Risk for Coronary Events **Evidence** that use of ACE-I and ARB are associated with lower mortality and/or cardiac events **#0058** Inappropriate antibiotic treatment for adults with acute bronchitis **Evidence** that antibiotics are not effective for acute bronchitis
Intermediate Clinical Outcome An intermediate outcome is a change in physiologic state that leads to a longer-term health outcome	Quantity, quality, and consistency of a body of evidence that the measured intermediate clinical outcome leads to desired health outcomes in the target population	**#0059** Hemoglobin A1c Management **Evidence** that hemoglobin A1c > 9 is associated with more complications
Health Outcome An outcome of care is a health state of a patient (or change in health status) resulting from healthcare – desirable or adverse In some situations, resource use measures may be considered proxies for a health state (e.g., hospitalization may represent a deterioration in health status)	Optimally, quantity, quality, and consistency for a body of evidence that the measured outcome (desirable or adverse) is influenced by at least one healthcare process or service. However, outcomes do not necessarily require evidence.	**#0230** Acute Myocardial Infarction 30-day Mortality Survival is a goal of seeking and providing treatment for AMI **Evidence** that healthcare processes/ interventions (aspirin, reperfusion) affect mortality/ survival **#0171** Acute care hospitalization (risk-adjusted) [of home care patients] Improvement or stabilization of condition to remain at home is a goal of seeking and providing home care services. **Evidence** that healthcare processes (e.g., medication reconciliation, care coordination) affect hospitalization of patients receiving home care services **#0140** Ventilator-associated pneumonia for ICU and high-risk nursery (HRN) patients Avoiding harm from treatment is a goal of when seeking and providing healthcare. **Evidence** that ventilator acquired pneumonia is affected by healthcare processes (e.g., ventilator

Type of Measure	Evidence	Example of Measure Type & Evidence to be Addressed
		bundle)
Special Considerations by Topic		
Patient experience with care	Evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information	**#0166** HCAHPS **Evidence** that patients/consumers value the aspects of care being measured (e.g., communication with doctors and nurses, responsiveness of hospital staff, pain control, communication about medicines, cleanliness and quiet of the hospital environment, and discharge information)
Efficiency Measures of efficiency combine the concepts of resource use and quality	**Efficiency Measured with combination of Quality measures and Resource Use measures** **Quality measure component** Evidence for the selected quality measure(s) as described in this table **Resource use measure component** Does not require clinical evidence as described in this table	Currently, there are no NQF-endorsed efficiency measures that combine quality and resource use Potential Measure: Diabetes quality measure(s) or composite used in conjunction with a measure of resource use per episode **Evidence** for diabetes quality measure(s) as described in this table

III. Recommendations for Evaluating Criterion 1c – Quantity, Quality, Consistency of Body of Evidence

The following recommendations and decision rules apply to evaluating evidence whether for initial endorsement, endorsement maintenance, or ad hoc review. The state of science may change over time, therefore at the time of review for endorsement maintenance, it also is appropriate to reexamine the evidence to assess whether new and innovative ways of organizing and providing care have evolved which achieve the same or better outcomes potentially at less cost.

- Evidence should be evaluated on *quantity*, *quality* of studies, *consistency* in direction, and magnitude of net benefit (benefits over harms) of a ***body of evidence*** on a scale of High, Moderate, or Low.
- The dimensions of *quantity*, *quality*, and *consistency* for a body of evidence apply to measures based on guidelines as well as those for which guidelines may not exist (e.g., care coordination or team functioning may not be based on guidelines, but often have bodies of evidence including non-clinical literature that should be systematically assessed)

- Measures without a clear description of the *quantity*, *quality*, and *consistency* of the supporting body of evidence or without any evidence should not pass criterion 1c and the threshold criterion of *Importance to Measure and Report*.
- Use of only selected individual studies from a body of evidence is not adequate to evaluate the evidence and should not pass criterion 1c and the threshold criterion of *Importance to Measure and Report*. This should be flagged in the measure submission form.
- Inconsistent and conflicting evidence should result in measures not passing both criterion 1c and the threshold criterion of *Importance to measure and report*.
- Expert opinion is acceptable evidence; it should be systematically assessed and fully described and will be evaluated as outlined in Table 4.

Table 4 provides guidance on how to evaluate each of the dimensions of *quantity*, *quality*, and *consistency* for a body of evidence. Table 5 provides recommended decision rules for using the ratings for all three dimensions to make a decision on whether a measure should pass the criterion 1c, the evidence to support the measure focus. High quality evidence usually requires multiple studies each with sufficient numbers of patients to give precise estimates, but occasionally a large and representative study can give high quality evidence. For example, one study (low quantity) that is a RCT with a large representative sample of patients (high quality) and substantial estimates of net benefit would pass the criterion, whereas, a body of evidence with low consistency of estimates of net benefits indicates a measure should not pass the criterion regardless of the ratings for quantity and quality of studies.

381 Table 4. Evaluation of Quantity, Quality, and Consistency of Body of Evidence for Criterion 1c –
382 evidence for the measure focus
383

	Quantity of Body of Evidence	**Quality** of Body of Evidence	**Consistency** of Results of Body of Evidence
Definition	Total number of studies (not articles or papers)	Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors* including: study design or flaws; directness/indirectness regarding the specific process or structure being measured, outcomes assessed, target population, comparisons; imprecision (wide confidence intervals due to few patients or events)	Stability in both the magnitude and direction of benefits and harms to patients (benefits over harms) across studies in the body of evidence
High	5+ studies**	Randomized controlled trials (RCTs) of direct evidence, with adequate size to obtain precise estimates of effect, and without serious flaws that introduce bias	Estimates of benefits and harms to patients are consistent in direction and similar in magnitude across studies in the body of evidence
Moderate	2-4 studies**	• Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect; or • RCTs without serious flaws that introduce bias, but with either indirect evidence, or imprecise estimate of effect	Estimates of benefits and harms to patients are consistent in direction, but differ in magnitude across studies in the body of evidence; **or** If only one study, the estimate of benefits greatly outweighs the estimate of potential harms **OR** For expert opinion that is systematically assessed, agreement that benefits to patients clearly outweigh potential harms
Low	0-1 studies**	• Expert opinion that is systematically assessed; or • RCTs with flaws that introduce bias; or • Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations	Differences in both magnitude and direction of benefits and harms to patients across studies in the body of evidence; or wide confidence intervals prevent estimating net benefit **OR** For expert opinion evidence that is systematically assessed: lack of agreement that benefits to patients clearly outweigh potential harms
Inadequate to Evaluate	No empirical evidence; **OR** only selected individual studies from a larger body of evidence	Expert opinion only and it was not systematically assessed	No assessment of magnitude and direction of benefits and harms to patients

*Study designs that affect certainty of confidence in estimates of effect include: Randomized controlled trials (RCT), which control for both observed and unobserved confounders, and non-RCTs (observational studies) with various levels of control for confounders.
Study flaws that may bias estimates of effect include: lack of allocation concealment; lack of blinding; large losses to follow-up; failure to adhere to intention to treat analysis; stopping early for benefit; failure to report important outcomes.
Imprecision with wide confidence intervals around estimates of effects can occur in studies involving few patients and few events.
Indirectness of evidence includes: indirect comparisons (e.g., two drugs compared to placebos rather than head-to head), differences between the population, intervention, outcome of interest, or comparator interventions and those included in the relevant studies. [14]
** The suggested number of studies for rating levels of quantity is considered a general guideline.

Table 5. Evaluation of Criterion 1c based on the quantity, quality and consistency of the body of evidence

Quantity of Body of Evidence	Quality of Body of Evidence	Consistency of Body of Evidence	Pass Criterion 1c
Moderate-High	Moderate-High	Moderate-High	Yes
Low	Moderate-High	Moderate-High	Yes, but only if judgment that additional research is unlikely to change conclusion that benefits to patients outweigh harms; otherwise, No
Moderate-High	Low	Moderate-High	Yes, but only if judgment that potential benefits to patients clearly outweigh potential harms; otherwise, No
Low-Mod-High	Low-Mod-High	**Low**	No
Low	Low	Low	No

IV. Recommendations for Selecting the Focus for Measure Development

Based on its discussion and recommendations regarding evidence to support the measure focus, the following recommendations address selecting a focus for measure development.

- For any topic area, measures based on the best evidence should be considered over measures based on lower quality evidence (e.g., expert opinion).
- There is a hierarchical preference for outcome measures (when possible) followed by process measures. Structural measures are appropriate primarily when there are very well established structure-process-outcome relationships; and when it is not feasible to directly measure the outcomes or processes. For process and structure measures, the focus of measurement should be on the aspect of care with the most direct evidence of a strong

relationship to the desired outcome. For example, evidence about effective medication to control blood pressure is direct evidence for the medication but only indirect evidence for the frequency of assessing blood pressure (see Figure 2). Assessment of blood pressure, although necessary, is not sufficient to achieving control. When there are multiple processes that affect a desired outcome, efforts should be made to include measures for all processes that have a strong relationship to the desired outcome.

V. Recommendations for Evaluating Importance to Measure and Report and the Other Subcriteria

Although the criterion *Importance to Measure and Report* has been a threshold, must-pass criterion, the weight of the individual subcriteria in making the determination of whether the criterion was met was not specified. The Task Force recommended that all three subcriteria must be met: High impact (1a), Opportunity for improvement (1b), and Evidence for the focus of measurement (1c) as noted above.

Generally, in measure submissions, high impact is easily demonstrated by alignment with a specific NPP goal or epidemiologic data (incidence, prevalence, resource use, consequences of quality problems). However, data on opportunity for improvement may be lacking (e.g., submitter states that performance is unknown), or it may not be specific to the focus of measurement, or only based on a sample from measure development and testing. When data are presented, reviewers sometimes question whether there is enough variation to justify importance to measure and report or how to judge overall poor performance. When a measure undergoes review for continued endorsement, one issue that sometimes arises is whether a measure is "topped out" meaning there are high levels of performance with little variation and therefore, little room for further improvement. The Task Force did not recommend specific quantitative thresholds for identifying conformance with the subcriteria of high impact (1a) and opportunity for improvement (1b).

- Threshold values for opportunity for improvement would be difficult to standardize. It depends on the size of the population at risk, effectiveness, and the consequences of the quality problem. For example, even modest variation would be sufficient justification for

some highly effective, potentially life-saving treatments (e.g., certain vaccinations) that are critical to the public health.

- At the time of review for continued endorsement, being "topped out" is not a reason in itself to remove endorsement for a measure; however, it may be a signal of some other problem with the measure (e.g. imprecise specification, overly broad exclusions). If a measure is an important and valid indicator of quality, it may still be justified to retain endorsement, as overall performance could deteriorate if not monitored. However, a "topped out" process measure might have endorsement withdrawn if there is an associated outcome measure.

Table 6. Evidence for Evaluating Importance to Measure and Report

Subcriterion	Evidence	Example	Pass the subcriterion?
High impact (1a)	Addresses a specific national health goal/priority identified by the Secretary of DHHS or the NPP **OR** Epidemiologic data – affects large numbers of patients and/or has a very substantial impact for smaller populations; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and patient/societal consequences of poor quality	#0140 Ventilator-associated pneumonia for ICU and high-risk nursery (HRN) patients **NPP goal: . . .** focus relentlessly on continually reducing and seeking to eliminate all healthcare-associated infections (HAIs) **Evidence** related to numbers of patients (e.g., 250,205 VAPs reported 35,969 (14.4%) were fatal; cost (e.g., total annual cost of VAP $2.5 billion)	**Subcriterion 1a** **Yes** – Demonstrated at least one of the aspects of high impact **No** – Did not demonstrate at least one of the aspects of high impact
Opportunity for improvement (1b)	Epidemiologic data; health services research – data demonstrating considerable variation, or overall poor performance, in the quality of care across providers and/or population groups (disparities in care)	#0432 Influenza Vaccination of Nursing Home/ Skilled Nursing Facility Residents **NPP goal:** All Americans will receive the most effective preventive services recommended by the U.S. Preventive Services Task Force **Evidence** that vaccination rates vary (e.g., 39% fail to reach the Healthy People 2010 objective of vaccinating at least 90% of nursing home residents)	**Subcriterion 1b** **Yes** – Demonstrated either variation or overall less than optimal performance **No** – Did not demonstrate either variation **or** overall less than optimal performance
Evidence for the focus of measurement (1c)	See Table 3	See Table 3	**Subcriterion 1c** See Table 4 and Table 5
			Pass Criterion, Importance to Measure and Report?
All 3 subcriteria (1a,1b,1c) must be met to pass the threshold criterion, *Importance to Measure and Report*			

VI. Recommendations for Modifications to the NQF Evaluation Criteria

As noted previously, the Task Force recommended that all three subcriteria be met to pass the threshold criterion of *Importance to Measure and Report*. The following redlined modifications to the criteria are based on the Task Force recommendations as reported above, as well as a few editorial changes.

1. Importance to measure and report: Extent to which the specific measure focus is evidence-based, important to making significant gains in health care quality ~~(safety, timeliness, effectiveness, efficiency, equity, patient-centeredness)~~ and improving health outcomes for a specific high impact aspect of healthcare where there is variation in or overall poor performance. ***Candidate measures must be judged to be important to measure and report*** *in order to be evaluated against the remaining criteria.*

1a. The measure focus addresses:

- a specific national health goal/priority identified by ~~NQF's~~ DHHS or the National Priorities Partnership convened by NQF;

  OR
- a demonstrated high impact aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

AND

1b. Demonstration of quality problems and opportunity for improvement, i.e., data (footnote 1) demonstrating considerable variation, or overall poor performance, in the quality of care across providers and/or population groups (disparities in care).

AND

1c. The measure focus is evidence-based as demonstrated by a systematic assessment of the quantity, quality, and consistency of the body of evidence (see Tables 3-5) and standardized grading of the body of evidence (footnote 3).

- ~~an~~ Health outcome (footnote 2): optimally, evidence that the measured outcome (desirable or adverse) is influenced by at least one healthcare process or service. However, outcomes do not necessarily require evidence. ~~(e.g., morbidity, mortality, function, health-related quality of life) that is relevant to, or associated with, a national health goal/priority, the condition, population, and/or care being addressed (2);~~

  ~~OR~~
- ~~if an intermediate outcome, process, structure, etc., there is **evidence (3)** that supports the specific measure focus as follows:~~
- Intermediate clinical outcome: evidence that the measured intermediate outcome (e.g., blood pressure, Hba1c) leads to desired outcomes in the target population ~~improved health/avoidance of harm or cost/benefit~~.
- Process (footnote 4): evidence that the measured ~~clinical or administrative~~ healthcare process leads to desired outcomes in the target population~~improved health/avoidance of harm and if the measure focus is on one step in a multi-step care process (4), it measures the step that has the greatest effect on improving the specified desired outcome(s)~~.
- Structure: evidence that the measured structure leads to desired health outcomes (including evidence for the link to effective care processes and the link from the care processes to desired health outcomes) ~~supports the consistent delivery of effective processes or access that lead to improved health/avoidance of harm or cost/benefit~~.
- Special Considerations by Topic of Measurement
  - Patient experience with care: – evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information~~an association exists between the measure of patient experience of health care and the outcomes, values and preferences of individuals/ the public~~.
  - ~~Access – evidence that an association exists between access to a health service and the outcomes of, or experience with, care.~~

- ~~o~~ Efficiency (footnote 5): evidence for the quality component as noted above~~demonstration of an association between the measured resource use and level of performance with respect to one or more of the other five IOM aims of quality~~.

Footnotes

[1] Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, or data from pilot testing or implementation of the proposed measure. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

[2] Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, ~~"never events"~~ serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

[3] The ~~strength of the body of evidence for the specific measure focus should be systematically assessed and rated (e.g.,~~ preferred systems for grading the evidence are the USPSTF ~~grading system~~ (grading definitions and methods) or GRADE~~)~~. ~~If the USPSTF grading system was not used, the grading system is explained including how it relates to the USPSTF grades or why it does not. However, evidence is not limited to quantitative studies and the best type of evidence depends upon the question being studied (e.g., randomized controlled trials appropriate for studying drug efficacy are not well suited for complex system changes). When qualitative studies are used, appropriate qualitative research criteria are used to judge the strength of the evidence.~~

[4] Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multi-step process, the step with the ~~greatest effect on~~ strongest evidence for the link to the desired outcome should be selected as the focus of measurement. ~~For example, although assessment of immunization status and recommending immunization are necessary steps, they are not sufficient to achieve the desired impact on health status – patients must be vaccinated to achieve immunity. This does not preclude consideration of measures of preventive screening interventions where there is a strong link with desired outcomes (e.g., mammography) or measures for multiple care processes that affect a single outcome.~~

[5] Measures of efficiency combine the concepts of resource use and quality ~~Efficiency of care is a measurement construct of cost of care or resource utilization associated with a specified level of quality of care. It is a measure of the relationship of the cost of care associated with a specific level of performance measured with respect to the other five IOM aims of quality. Efficiency might be thought of as a ratio, with quality as the numerator and cost as the denominator. As such, efficiency is directly proportional to quality, and inversely proportional to cost.~~ (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; ~~based on~~ AQA Principles of Efficiency Measures).

VII. Recommendations for Modifications to the Measure Submission

The information requested on NQF's measure submission form is consistent with those identified in a 2009 collaborative effort undertaken with CMS, The Joint Commission, NCQA, and PCPI to identify common data fields. AHRQ participated, but maintained its own data items for the National Quality Measures Clearinghouse.

The Task Force suggested the following modifications to the information requested on the NQF measure submission form. The intent is full transparency about the supporting evidence for the submitted measure. This will facilitate understanding of the adequacy of the evidence presented (selected evidence vs. a body of evidence) and the developer's representation of the quality of the evidence. Currently, evidence graded using the USPSTF or GRADE systems may not be

available, however, an accurate description of the evidence and any grading system used should still be expected. The following items pertain to the recommendations related to evidence (1c) under *Importance to Measure and Report*.

Table 7. Current and Proposed Information Requested on Measure Submission

Current Measure Submission (4.1) Items	Proposed Measure Submission Items
	Introduction *Importance to Measure and Report* is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria (1a, 1b, and 1c) must be met in order to pass this criterion. The following items request the information the committees will need to evaluate whether the criterion is met.
1c.1. Relationship to Outcomes *(For non-outcome measures, briefly describe the relationship to desired outcome. For outcomes, describe why it is relevant to the target population.)* **1c.2. Type of Evidence** *(Check all that apply)* Cohort study Observational study Evidence-based guideline Randomized controlled trial Expert opinion Systematic synthesis of research Meta-analysis Other: **1c.4. Summary of Evidence** *(For non-outcome measures, provide evidence of relationship to desired outcome. For outcomes, summarize any evidence that healthcare services/care processes influence the outcome.)* **1c.5. Rating of Strength/Quality of Evidence** *(Also provide narrative description of the rating and by whom)* **1c.6. Method for Rating Evidence** **1c.7. Summary of Controversy/Contradictory Evidence** **1c.8. Citations for Evidence** *(Other than guidelines)* **1c.9. Quote the Specific Guideline Recommendation** *(Including guideline number and/or page number)*	**1c.1. Structure-Process-Outcome Relationship** *(Briefly state the measured structure, process, or outcome and the links and direction between: a) the measured process and desired outcome; b) the measured outcome and processes that influence the outcome; or c) the measured structure and effective processes and desired outcome.)* **1c.2. Source of Evidence** Clinical practice guideline Systematic review of body of evidence (other than within guideline development) Selected individual studies (rather than entire body of evidence) Other **1c.4. Summary of Body of Evidence** Quantity of Studies in Body of Evidence (*total number of studies, not articles*): Quality of Body of Evidence (*Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors including: study design/ flaws; directness/indirectness regarding the specific process/structure being measured, outcomes assessed, target population, comparisons; imprecision (wide confidence intervals due to few patients or events*): Directness to focus of measurement & target population in proposed measure: Consistency of Results across Studies: Net Benefit (B*enefits over harms*) Benefit/outcome – estimate of effect Harms addressed – estimate of effect **1c.5. Grading of Strength/Quality of Body of Evidence** Has the **body of evidence** been graded? Yes No If graded: By whom (*describe the entity that graded the evidence,*

Current Measure Submission (4.1) Items	Proposed Measure Submission Items
1c.10. Clinical Practice Guideline Citation **1c.11. National Guideline Clearinghouse or Other URL** **1c.12. Rating Strength of Recommendation** *(Also provide narrative description of the rating and by whom)* **1c.13. Method for Rating Strength of Recommendation** *(If different from USPSTF system, also describe rating and how it relates to USPSTF)* **1c.14. Rationale for Using This Guideline Over Others**	*including balance of representation and any disclosures regarding bias*) Grade Assigned to the Evidence: Based on the NQF descriptions for rating the evidence, what was your assessment of the body of evidence (rate each as High, Moderate, or Low) Quantity: Quality: Consistency: **1c.6. System for Grading Evidence:** USPSTF GRADE Other (*provide description of grading scale with definitions*) **1c.7. Summary of Controversy/Contradictory Evidence** **1c.8. Citations for Evidence** *(Other than guidelines)* **1c.9. Quote Verbatim the Specific Guideline Recommendation** *(Including guideline number and/or page number)* **1c.10. Clinical Practice Guideline Citation** **1c.11. National Guideline Clearinghouse or Other URL for the cited guideline** **1c.12. Grading of Strength of Recommendation** Has the **recommendation** been graded? Yes No If graded: By whom (*describe the entity that graded the evidence, including balance of representation and any disclosures regarding bias*) Grade Assigned to the Recommendation: **1c.13. System for Grading Strength of Recommendation:** USPSTF GRADE Other (*provide description of grading scale with definitions*) **1c.14. Rationale for Using This Guideline Over Others**
Descriptive Information **De.4. National Priority Partnership priority area** *(Select the most relevant)* Patient and family engagement Population health Safety Care coordination Palliative and end of life care Overuse **De.5. IOM Quality Domain** *(Select the most relevant)* Effectiveness Efficiency	**Descriptive Information – no change** **De.4. National Priority Partnership priority area** *(Select the most relevant)* Patient and family engagement Population health Safety Care coordination Palliative and end of life care Overuse **De.5. IOM Quality Domain** *(Select the most relevant)* Effectiveness Efficiency

Current Measure Submission (4.1) Items	Proposed Measure Submission Items
Equity	Equity
Patient-centered	Patient-centered
Safety	Safety
Timeliness	Timeliness
De.6. Consumer Care Need *(Select the most relevant)*	**De.6. Consumer Care Need** *(Select the most relevant)*
Getting better	Getting better
Living with illness	Living with illness
Staying healthy	Staying healthy

VIII. Recommendations for Evidence Required for Practices Considered for NQF Endorsement

NQF also endorses practices such as safe practices, care coordination practices, and substance use treatment practices. The criteria for practices include evidence of effectiveness.

The Task Force recommends that the same evidence requirements as indicated for process measures (Tables 3, 4, 5) be applied to practices considered for NQF endorsement.

Table 8. Evidence to Support a Practice

Evidence to Support a Practice	Example of Practice & Evidence to be Addressed
Quantity, quality, and consistency of a body of evidence that the measured healthcare process leads to desired health outcomes in the target population	**Safe Practice 16** Safe Adoption of Computerized Prescriber Order Entry **Evidence** that computerized order entry systems are associated with lower medication errors and adverse events

Modifications to Practice Evaluation Criteria

Evidence of Effectiveness. A practice is evidence-based as demonstrated by a systematic assessment of the quantity, quality, and consistency of the body of evidence (see Tables 3-5) and standardized grading of the body of evidence (footnote).~~There must be clear evidence that the practice would be effective in reducing patient safety events. Such evidence may take various forms, including the following:~~

~~o Research studies showing a direct connection between improved clinical outcomes (e.g., reduced mortality or morbidity) and the practice;~~

~~o experiential data (including broad expert agreement, widespread opinion, or professional consensus) showing the practice is "obviously beneficial" or self-evident (i.e., the practice absolutely constrains a potential problem or forces an improvement to occur, reduces reliance on memory, standardizes equipment or process steps, or promotes teamwork); or~~

~~o Research findings or experiential data~~ Evidence from non-healthcare industries that should be substantially transferable to healthcare (e.g., safety practices of repeat-back of verbal orders or standardizing abbreviations) also may be considered.

Footnote:
The preferred systems for grading the evidence are the USPSTF (grading definitions and methods) or GRADE.

Consequences of Measurement

Consequences of measurement are not the same as the consequences of the measured structure or process, i.e., the benefits or harms to the patient related to the specific topic of measurement. Currently, unintended consequences of measurement are addressed under feasibility.

> **4d.** Susceptibility to inaccuracies, errors, or unintended consequences of measurement and the ability to audit the data items to detect such problems are identified.

The Task Force identified that actual vs. theoretical consequences to measurement are most likely to arise after implementation and should be addressed at the time of review for continued endorsement. For example, a measure of timing of antibiotic administration in patients with pneumonia may result in some patients receiving antibiotics before the diagnosis of pneumonia is confirmed by x-ray. The Task Force did not recommend moving subcriterion 4d under *Importance to Measure and Report*, but might it could be considered a threat to validity.

REFERENCES

1. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. *Int J Qual Health Care*. 2004;16(1):9-18.
2. Spertus JA, Eagle KA, Krumholz HM et al. American College of Cardiology and American Heart Association methodology for the selection and creation of performance measures for quantifying the quality of cardiovascular care. *Circulation*. 2005;111(13):1703-1712.
3. Tricoci P, Allen JM, Kramer JM et al. Scientific evidence underlying the ACC/AHA clinical practice guidelines. *JAMA*. 2009;301(8):831-841.
4. Physician Consortium for Performance Improvement. Physician Consortium for Performance Improvement® (PCPI) Position Statement - The Evidence Base Required for Measures Development. *American Medical Association* 6-26-2009;1-18. Last accessed March 2010.
5. Grilli R, Magrini N, Penna A et al. Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet*. 2000;355(9198):103-106.

6. Shiffman RN, Shekelle P, Overhage JM et al. Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization. *Ann Intern Med*. 2003;139(6):493-498.
7. The AGREE Collaboration. *Appraisal of Guidelines for Research and Evaluation AGREE Instrument*. 2001. Available at http://www.agreecollaboration.org/instrument/. Last accessed February 2010.
8. The AGREE Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. *Quality & safety in health care*. 2003;12(1):18-23.
9. Atkins D, Eccles M, Flottorp S et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. *BMC Health Serv Res*. 2004;4(1):38.
10. Atkins D, Best D, Briss PA et al. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328(7454):1490-1494.
11. Guyatt GH, Oxman AD, Vist GE et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.
12. Guyatt GH, Oxman AD, Kunz R et al. Incorporating considerations of resources use into grading recommendations. *BMJ*. 2008;336(7654):1170-1173.
13. Guyatt GH, Oxman AD, Kunz R et al. Going from evidence to recommendations. *BMJ*. 2008;336(7652):1049-1051.
14. Guyatt GH, Oxman AD, Kunz R et al. What is "quality of evidence" and why is it important to clinicians? *BMJ*. 2008;336(7651):995-998.
15. Guyatt GH, Oxman AD, Vist GE et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.
16. Harris RP, Helfand M, Woolf SH et al. Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med*. 2001;20(3 Suppl):21-35.
17. Sawaya GF, Guirguis-Blake J, LeFevre M et al. Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Ann Intern Med*. 2007;147(12):871-875.

18. Owens DK, Lohr KN, Atkins D et al. Grading the strength of a body of evidence when comparing medical interventions-Agency for Healthcare Research and Quality and the Effective Health Care Program. *J Clin Epidemiol*. 2009.
19. Liberati A, Altman DG, Tetzlaff J et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med*. 2009;151(4):W65-W94.
20. Donabedian A. *An Introduction to Quality Assurance in Health Care*. New York, NY: Oxford University Press; 2003.
21. Donabedian A. The role of outcomes in quality assessment and assurance. *Quality Review Bulletin*. 1992;18(11):356-360.

TASK FORCE MEMBERS

David Shahian, MD (chair)
Center for Quality and Safety and Department of Surgery,
Massachusetts General Hospital
Professor of Surgery, Harvard Medical School

Kristine Martin Anderson, MBA
Senior Vice President, Booz Allen Hamilton, Rockville, MD
Consensus Standards Approval Committee (CSAC) member

David Atkins MD, MPH
Director of Quality Enhancement Research Initiative (QUERI),
Department of Veterans Affairs, Health Services Research & Development Service

Arthur Levin, MPH
Director, Center for Medical Consumers, New York, NY
Consensus Standards Approval Committee (CSAC) member

Mary Naylor, PhD, RN
Marian S. Ware Professor in Gerontology
University of Pennsylvania School of Nursing
Board member

Greg Pawlson, MD, MPH
Executive Vice President, National Committee for Quality Assurance (NCQA)

Eric Schneider, MD, MSc, FACP
Senior Scientist and Director, RAND Boston
Associate Professor, Division of General Medicine and Primary Care
Brigham and Women's Hospital and
Department of Health Policy and Management
Harvard School of Public Health

APPENDIX

NATIONAL QUALITY FORUM

Current Measure Evaluation Criteria
December 2009

Conditions for Consideration

Four conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards:

A. The measure is in the public domain or an intellectual property agreement is signed.

B. The measure owner/steward verifies there is an identified responsible entity and process to maintain and update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every 3 years.

C. The intended use of the measure includes both public reporting and quality improvement.

D. The requested measure submission information is complete. Generally, measures should be fully developed and tested so that all the evaluation criteria have been addressed and information needed to evaluate the measure is provided. Measures that have not been tested are only potentially eligible for a time-limited endorsement and in that case, measure owners must verify that testing will be completed within ~~24~~ 12 months of endorsement.

Criteria for Evaluation

If all four conditions for consideration are met, candidate measures are evaluated for their suitability based on four sets of standardized criteria: importance to measure and report, scientific acceptability of measure properties, usability, and feasibility. Not all acceptable measures will be strong—or equally strong—among each set of criteria. The assessment of each criterion is a matter of degree; however, all measures must be judged to have met the first criterion, importance to measure and report, in order to be evaluated against the remaining criteria.

1. Importance to measure and report: Extent to which the specific measure focus is important to making significant gains in health care quality (safety, timeliness, effectiveness, efficiency, equity, patient-centeredness) and improving health outcomes for a specific high impact aspect of healthcare where there is variation in or overall poor performance. ***Candidate measures must be judged to be important to measure and report*** *in order to be evaluated against the remaining criteria.*

1a. The measure focus addresses:

- a specific national health goal/priority identified by NQF's National Priorities Partners; OR
- a demonstrated high impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use (current and/or future), severity of illness, and patient/societal consequences of poor quality).

1b. Demonstration of quality problems and opportunity for improvement, i.e., data[1] demonstrating considerable variation, or overall poor performance, in the quality of care across providers and/or population groups (disparities in care).

1c. The measure focus is:

- an outcome (e.g., morbidity, mortality, function, health-related quality of life) that is relevant to, or

[1] Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, measure data from pilot testing or implementation. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

associated with, a national health goal/priority, the condition, population, and/or care being addressed[2];
OR

- if an intermediate outcome, process, structure, etc., there is **evidence**[3] that supports the specific measure focus as follows:
  - Intermediate outcome – evidence that the measured intermediate outcome (e.g., blood pressure, Hba1c) leads to improved health/avoidance of harm or cost/benefit.
  - Process – evidence that the measured clinical or administrative process leads to improved health/avoidance of harm and
    if the measure focus is on one step in a multi-step care process[4], it measures the step that has the greatest effect on improving the specified desired outcome(s).
  - Structure – evidence that the measured structure supports the consistent delivery of effective processes or access that lead to improved health/avoidance of harm or cost/benefit.
  - Patient experience – evidence that an association exists between the measure of patient experience of health care and the outcomes, values and preferences of individuals/ the public.
  - Access – evidence that an association exists between access to a health service and the outcomes of, or experience with, care.
  - Efficiency[5] – demonstration of an association between the measured resource use and level of performance with respect to one or more of the other five IOM aims of quality.

If not important to measure and report, STOP.

2. Scientific acceptability of the measure properties: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

2a. The measure is well defined and precisely specified[6] so that it can be implemented consistently within and across organizations and allow for comparability. The required data elements are of high quality as defined by NQF's Health Information Technology Expert Panel (HITEP)[7] .

---

[2] Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, "never events" that are compared to zero are appropriate outcomes for public reporting and quality improvement.

[3] The strength of the body of evidence for the specific measure focus should be systematically assessed and rated (e.g., USPSTF grading system – grade definitions and methods). If the USPSTF grading system was not used, the grading system is explained including how it relates to the USPSTF grades or why it does not. However, evidence is not limited to quantitative studies and the best type of evidence depends upon the question being studied (e.g., randomized controlled trials appropriate for studying drug efficacy are not well suited for complex system changes). When qualitative studies are used, appropriate qualitative research criteria are used to judge the strength of the evidence.

[4] Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multi-step process, the step with the greatest effect on the desired outcome should be selected as the focus of measurement. For example, although assessment of immunization status and recommending immunization are necessary steps, they are not sufficient to achieve the desired impact on health status – patients must be vaccinated to achieve immunity. This does not preclude consideration of measures of preventive screening interventions where there is a strong link with desired outcomes (e.g., mammography) or measures for multiple care processes that affect a single outcome.

[5] Efficiency of care is a measurement construct of cost of care or resource utilization associated with a specified level of quality of care. It is a measure of the relationship of the cost of care associated with a specific level of performance measured with respect to the other five IOM aims of quality. Efficiency might be thought of as a ratio, with quality as the numerator and cost as the denominator. As such, efficiency is directly proportional to quality, and inversely proportional to cost. (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; based on AQA Principles of Efficiency Measures).

[6] Measure specifications include the target population (e.g., denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (e.g., numerator),

2b. Reliability testing[8] demonstrates the measure results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period.

2c. Validity testing[9] demonstrates that the measure reflects the quality of care provided, adequately distinguishing good and poor quality. If face validity is the only validity addressed, it is systematically assessed.

2d. Clinically necessary measure exclusions are identified and must be:

- supported by evidence[10] of sufficient frequency of occurrence so that results are distorted without the exclusion;

AND

- a clinically appropriate exception (e.g., contraindication) to eligibility for the measure focus[11];

AND

- precisely defined and specified:
  - if there is substantial variability in exclusions across providers, the measure is specified so that exclusions are computable and the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion);
  - if patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that it strongly impacts performance on the measure and the measure must be specified so that the information about patient preference and the effect on the measure is transparent[12] (e.g., numerator category computed separately, denominator exclusion category computed separately).

2e. For outcome measures and other measures (e.g., resource use) when indicated:

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are present at start of care[11,13]

---

measurement time window, exclusions, risk adjustment, definitions, data elements, data source and instructions, sampling, scoring/computation.

[7] The HITEP criteria for high quality data include: a) data captured from an authoritative/accurate source; b) data are coded using recognized data standards; c) method of capturing data electronically fits the workflow of the authoritative source; d) data are available in EHRs; and e) data are auditable. NQF. *Health Information Technology Expert Panel Report: Recommended Common Data Types and Prioritized Performance Measures for Electronic Healthcare Information Systems*. Washington, DC: NQF; 2008.

[8] Examples of reliability testing include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing may address the data items or final measure score.

[9] Examples of validity testing include, but are not limited to: determining if measure scores adequately distinguish between providers known to have good or poor quality assessed by another valid method; correlation of measure scores with another valid indicator of quality for the specific topic; ability of measure scores to predict scores on some other related valid measure; content validity for multi-item scales/tests. Face validity is a subjective assessment by experts of whether the measure reflects the quality of care (e.g., whether the proportion of patients with BP < 140/90 is a marker of quality). If face validity is the only validity addressed, it is systematically assessed (e.g., ratings by relevant stakeholders) and the measure is judged to represent quality care for the specific topic and that the measure focus is the most important aspect of quality for the specific topic.

[10] Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, sensitivity analyses with and without the exclusion, and variability of exclusions across providers.

[11] Risk factors that influence outcomes should not be specified as exclusions.

[12] Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

[13] Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care such as race, socioeconomic status, gender (e.g., poorer treatment outcomes of

OR

- rationale/data support no risk adjustment.

2f. Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful[14] differences in performance.

2g. If multiple data sources/methods are allowed, there is demonstration they produce comparable results.

2h. If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender);
OR
rationale/data justifies why stratification is not necessary or not feasible.

3. Usability: Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

3a. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audience(s) for both public reporting (e.g., focus group, cognitive testing) and informing quality improvement (e.g., quality improvement initiatives)[15]. An important outcome that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.

3b. The measure specifications are harmonized[16] with other measures, and are applicable to multiple levels and settings.

3c. Review of existing endorsed measures and measure sets demonstrates that the measure provides a distinctive or additive value to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of healthcare).

4. Feasibility: Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement.

4a. For clinical measures, required data elements are routinely generated concurrent with and as a

---

African American men with prostate cancer, inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than adjusting out differences.

[14] With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74% v. 75%) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall poor performance may not demonstrate much variability across providers.

[15] Public reporting and quality improvement are not limited to provider-level measures – community and population measures also are relevant for reporting and improvement.

[16] Measure harmonization refers to the standardization of specifications for similar measures on the same topic (e.g., ***influenza immunization*** of patients in hospitals or nursing homes), or related measures for the same target population (e.g., eye exam and HbA1c for ***patients with diabetes***), or definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless differences are dictated by the evidence. The dimensions of harmonization can include numerator, denominator, exclusions, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.

byproduct of care processes during care delivery.

4b. The required data elements are available in electronic sources. If the required data are not in existing electronic sources, a credible, near-term path to electronic collection by most providers is specified and clinical data elements are specified for transition to the electronic health record.

4c. Exclusions should not require additional data sources beyond what is required for scoring the measure (e.g., numerator and denominator) unless justified as supporting measure validity.

4d. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified.

4e. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality[17], etc.) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

If a measure meets the above criteria <u>and</u> there are competing measures (either endorsed measures, or other new submissions that also meet the criteria), compare measures on: Scientific acceptability of measure properties, Usability, and Feasibility to determine best-in-class.

5. Demonstration that the measure is superior to competing measures – new submissions and/or endorsed measures (e.g., is a more valid or efficient way to measure).

[17] All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.

Current Evaluation Criteria for Practices

Specificity. The practice must be a clearly and precisely defined process or manner of providing a healthcare service. All candidate safe practices were screened according to this threshold criterion. Candidate safe practices that met the threshold criterion of specificity were then rated against four additional criteria relating to the likelihood of the practice improving patient safety.

Benefit. If the practice were more widely utilized, it would save lives endangered by healthcare delivery, reduce disability or other morbidity, or reduce the likelihood of a serious reportable event (e.g., an effective practice already in near universal use would lead to little new benefit to patients by being designated a safe practice).

Evidence of Effectiveness. There must be clear evidence that the practice would be effective in reducing patient safety events. Such evidence may take various forms, including the following:
o Research studies showing a direct connection between improved clinical outcomes (e.g., reduced mortality or morbidity) and the practice;
o experiential data (including broad expert agreement, widespread opinion, or professional consensus) showing the practice is "obviously beneficial" or self-evident (i.e., the practice absolutely constrains a potential problem or forces an improvement to occur, reduces reliance on memory, standardizes equipment or process steps, or promotes teamwork); or
o Research findings or experiential data from non-healthcare industries that should be substantially transferable to healthcare (e.g., repeat-back of verbal orders or standardizing abbreviations).

Generalizability. The safe practice must be able to be utilized in multiple applicable clinical care settings (e.g., a variety of inpatient and/or outpatient settings) and/or for multiple types of patients.

Readiness. The necessary technology and appropriately skilled staff must be available to most healthcare organizations.

US Preventive Services Task Force System for Grading Evidence and Recommendations

The following information was obtained from AHRQ websites describing the grade definitions and methods.

What the Grades Mean and Suggestions for Practice

The USPSTF updated its definitions of the grades it assigns to recommendations and now includes "suggestions for practice" associated with each grade. The USPSTF has also defined levels of certainty regarding net benefit. These definitions apply to USPSTF recommendations voted on after May 2007.

Grade	Definition	Suggestions for Practice
A	The USPSTF recommends the service. There is high certainty that the net benefit is substantial.	Offer or provide this service.
B	The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.	Offer or provide this service.
C	The USPSTF recommends against routinely providing the service. There may be considerations that support providing the service in an individual patient. There is at least moderate certainty that the net benefit is small.	Offer or provide this service only if other considerations support the offering or providing the service in an individual patient.
D	The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits.	Discourage the use of this service.
I Statement	The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.	Read the clinical considerations section of USPSTF Recommendation Statement. If the service is offered, patients should understand the uncertainty about the balance of benefits and harms.

Levels of Certainty Regarding Net Benefit

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

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

U.S. Preventive Services Task Force Recommendation Grid*

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

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

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

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

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