NATIONAL QUALITY FORUM

Patient Reported Outcomes Workshop #2
September 11-12

Location: 1030 15th Street NW, 9th Floor Conference Center
Wi-Fi Network: Logon/user: guest; Password: NQFguest

Audience/General Registration number: (877) 303-9138 (both days)

Conference ID **Day 1**: 20945526 Conference ID **Day 2**: 21017521

Webinar Link **Day 1**: http://nqf.commpartners.com/se/NQFLogin/ Webinar Meeting ID Day 1: 323476
Webinar Link **Day 2**: http://nqf.commpartners.com/se/NQFLogin/ Webinar Meeting ID Day 2: 391249

Meeting Objectives:

1. Discuss the major methodological issues related to reliability and validity when aggregating PROM data into a performance measure;
2. Identify unique considerations in relation to the NQF endorsement criteria for PRO-based performance measures (PRO-PM) (as compared to other quality outcome performance measures);
3. Lay out the critical path from PROM to PRO-PM endorsed by NQF for use in accountability and performance improvement.

Terms

Patient-reported outcome (PRO): The concept of any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else.[1]
PRO domains included in this project encompass functional status/health-related quality of life; symptom/symptom burden; experience with care; health-related behaviors

PRO measure (PROM): Instrument, scale, or single-item measure used to assess the PRO concept as perceived by the patient, obtained by directly asking the patient to self-report (e.g., PHQ-9).

Performance measure: Numeric quantification of healthcare quality for a designated accountable healthcare entity, such as hospital, health plan, nursing home, clinician, etc.

PRO-based performance measure (PRO-PM): A performance measure that is based on PROM data aggregated for an accountable healthcare entity (e.g., percentage of patients in an accountable care organization whose depression score as measured by the PHQ-9 improved).

[1] U.S. FOOD AND DRUG ADMINISTRATION. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Federal Register 2009;74(35):65132-133. Available here.

AGENDA

Day 1 – September 11

8:30-9:00 Continental Breakfast & Networking (provided for Expert Panel/Authors)

9:00-9:30 Welcome & Setting the Stage
Patricia Brennan*, University of Wisconsin-Madison &* Joyce Dubow*, AARP, Co-chairs*

- Context
  - NQF endorses PRO-PMs, not the PROMs
  - NQF endorses PRO-PMs for use in accountability applications such as public reporting and payment
  - NQF evaluates suitability for endorsement based on a set of evaluation criteria
- Product of the workshop: Pathway from PROM to NQF-endorsed PRO-PM for use in accountability applications, including identification of unique considerations in relation to the NQF endorsement criteria taking into account the key methodological issues
  - Draft pathway from PROM to PRO-PM
    Drawing from the commissioned papers and groundwork from 1st workshop, a "straw man" pathway (represented in a flow schematic) serves as a starting point and will be refined on Day 2.

Expert Panel and Audience Engagement

9:35-11:05 **Lessons from the field – using PRO-PMs for accountability (public reporting, payment)**
Our international experts will be joining via webinar
Moderator: Greg Pawlson

Panel: England**–**David Nuttall, *Branch Head - Choice, AQP & PROMs, Strategy, Finance and NHS Directorate, Department of Health*; Medicare Advantage**–**Elizabeth Goldstein, *Director Division of Consumer Assessment and Plan Performance, Centers for Medicare and Medicaid Services* ; Sweden**–** Stefan Larsson, *Senior Partner & Managing Director Stockholm Office, Boston Consulting* Group

- How are the PRO-PMs being used (e.g., public reporting, payment, policy) and what are their impact?
- What PROMs were implemented for performance measurement and accountability and what were the key characteristics or considerations used for selection?
- How were the PROM data aggregated into performance measures (e.g., average/median amount of change; percentage who improve/reach benchmark/ have meaningful change)?
- How were reliability and validity of the PRO-PM demonstrated (beyond reliability and validity of the PROM)?
- How were threats to validity addressed (e.g., risk adjustment, response rate, missing data)?

Expert Panel and Audience Engagement

11:05-11:20 **BREAK**

11:25-12:45 **Recap of Key Characteristics for Selecting PROMs for Use in Performance Measurement**
Moderator: Karen Adams, *NQF*

- Key Characteristics
  Prior to the workshop, the Expert Panel reviewed potential additions to the characteristics identified in the 1st paper.

Karen Pace, *NQF* Overview of related NQF endorsement criteria: evidence; usability and use; feasibility

Panel: Elizabeth Mort, *Massachusetts General Hospital*; Laurie Burke, *Food and Drug Administration*; Jennifer Eames-Huff, *Pacific Business Group on Health*

- Psychometric properties (Table from 1st paper)
- Actionability – responsiveness to healthcare intervention; facilitates buy-in from healthcare providers/clinicians
  What evidence is suggested – that clinical interventions affect the PRO, or that the patient/ person is the best source for assessing the PRO?
- Meaningfulness to patient/person and engagement in selecting PROMs
  How is patient/person engagement in selecting PROM i achieved and demonstrated?
- Implementable –PROM first used successfully in clinical care, not just collected for performance measurement; translation from research to practice
  What is needed to demonstrate that a PROM is implementable?

Expert Panel and Audience Engagement

12:45-1:30 **LUNCH** (provided for Expert Panel/Authors)

1:30-2:50 **Methods that Contribute to Trust– Demonstrating Reliability of PRO-PMs**

This discussion is based on the premise that the performance measure is based on a PROM that meets selection characteristics identified in Workshop#1.

Moderator: Karen Pace

- Overview of NQF endorsement criteria on reliability and differentiation between PROM & PRO-PM

Laura Smith, *RTI International* , commissioned paper author tees-up key issues and best practices or strengths/weaknesses of approaches

Reactor Panel: Lewis Kazis, *Boston University School of Public Health*; Lori Frank, *Patient Centered Outcomes Research Institute*; Jack Fowler, *Informed Medical Decisions Foundation*

- What impact does poor reliability of the PRO-PM score have on validity of the PRO-PM as an indicator of quality for the accountable healthcare entity?
- Are there any differences or unique considerations for demonstrating and evaluating the reliability of a PRO-PM (as compared to other quality performance measures)? Is reliability of the PRO-PM score needed in addition to reliability of the PROM?
- What methods for reliability testing would support the demonstration of reliability of the PRO-PM scores (e.g., signal to noise)?
- What are the implications of various approaches to aggregating PROM data (e.g., average/median amount of change; percentage who improve/reach benchmark/ have meaningful change) on reliability of the PRO-PM score?

Expert Panel and Audience Engagement

2:50-3:05 **BREAK**

3:10-4:30 **Methods that Contribute to Trust– Demonstrating Validity of PRO-PMs, Part 1**

This discussion is based on the premise that the performance measure is based on a PROM that meets selection characteristics identified in Workshop#1; and will focus on the construction of a performance measure and validity of inferences about quality of care.

Moderator: Karen Pace

- Overview of NQF endorsement criteria on validity and differentiation between PROM & PRO-PM

Anne Deutsch, *RTI International* & Barbara Gage, *Brookings Institution*, commissioned paper authors tee-up key issues and best practices or strengths/weaknesses of approaches

Reactor Panel: Stephan Fihn, *Veterans Health Administration*; Albert Wu, *Johns Hopkins Health System*

- What are the implications of various approaches to aggregating PROM data (e.g., average/median amount of change; percentage who improve/reach benchmark/ have meaningful change) on:
  - the validity of conclusions about quality; and
  - the ability to discriminate performance among accountable entities?
- What methods for validity testing would support the demonstration of validity of the performance measure score for making conclusions about quality of care?
- Are there any differences or unique considerations for demonstrating and evaluating the validity of PRO-PMs (as compared to other quality performance measures)?
- Is validity of the performance score as an indicator of quality needed in addition to validity of PROM?

Expert Panel and Audience Engagement

4:30-5:00 **Identification of Unique Considerations Related to NQF Endorsement of PRO-PMs**

Each table asked to identify unique considerations (as compared to other quality performance measures) for evaluating PRO-PMs as suitable for NQF endorsement

5:00 **Adjourn for the Day**

Day 2 – September 12

8:00-8:30 Continental Breakfast & Networking (provided for Expert Panel/Authors)

8:30-8:40 **Intro to Day 2**
Joyce Dubow

8:45-10:05 **Methods that Contribute to Trust – Demonstrating Validity of PRO-PMs, Part 2**
Discussion is based on the premise that the performance measure is based on a PROM that meets selection characteristics identified in Workshop#1; & will focus on real-world issues when implementing PRO-PMs that present challenges to inferences about quality of care.
Moderator: Karen Pace

- Overview of NQF endorsement criteria on threats to validity of conclusions about quality and differentiation between PROM & PRO-PM

Anne Deutsch tees-up key issues and best practices or strengths/ weaknesses of approaches to aggregating PROM data and specifying PRO-PMs

Reactor Panel: Kenneth Ottenbacher, *The University of Texas Medical Branch at Galveston*; Robert Weech-Maldonado, *University of Alabama at Birmingham*

- Are there any differences or unique considerations for risk adjustment of a PRO-PM (as compared to other quality outcome performance measures)?
- What are the implications of exclusions, incomplete/missing data, and response rate/bias on validity of the performance measure and the testing needed to assess impact on validity?
- What are the implications of using proxies on the validity of the performance measure and the testing needed to assess impact on validity?
- What are the implications of specifying more than one PROM (i.e., instrument/scale) in a performance measure and the testing needed to assess impact on validity?

Expert Panel and Audience Engagement

10:05-10:30 **Identification of Unique Consideration Related to NQF Endorsement of PRO-PMs**
Each table asked to identify unique considerations (as compared to other quality outcome performance measures) for evaluating PRO-PMs as suitable for NQF endorsement

10:30-10:45 **BREAK**

10:50-12:05 **Revisit pathway from individual-level PROM to NQF-endorsed PRO-PM**
Moderator: Karen Adams
Panel: Ethan Basch, *Memorial Sloan-Kettering Cancer Center,* (conceptual basis- steps 1-4); Jim Bellows, *Kaiser Permanente Care Management Institute*, (process performance measure-steps 5-9); Eleanor Perfetto, *Pfizer* (outcome performance measure-steps 10-13)

- Are all the steps in the pathway identified and in the correct order?
- Should performance measures begin with process measures (vs. outcome measures)?
- Is there flexibility in specifying multiple PROMs for process measures? Outcome measures?
- Along the various steps of the pathway identify:
    - Any unique considerations for endorsement of PRO-PMs (as compared to other quality outcome performance measures)
    - Guiding principles

Expert Panel and Audience Engagement

12:10-12:55 **LUNCH** (provided for Expert Panel/Authors)

1:00-1:45 **Future Directions**
Moderator: Patti Brennan
Expert Panel and Audience Engagement

- How do you see use of PROMs and PRO-PMs evolving in the future? (e.g., multiple individual-level PROs calibrated to a standard scale; use in composite measures)
- Do the foundations being built now (e.g., IT, evaluation criteria, and pathway) support the future?

1:50-2:00 **Closing Remarks and Next Steps**
Patti Brennan

Patient-Reported Outcomes in Performance Measurement

Commissioned Paper on
PRO-Based Performance Measures for Healthcare Accountable Entities

Draft # 1, September 4, 2012
Prepared for NQF PRO Workshop #2 – September 11-12, 2012

Anne Deutsch, RN, PhD, CRRN;[1] Barbara Gage, PhD;[2] Laura Smith, PhD;[1]
Cynthia Kelleher, MPH, MBA[1]

[1]RTI International, [2]Brookings Institution

Introduction

The National Quality Forum (NQF) has commissioned two papers on the use of Patient-Reported Outcome Measures (PROMs) in performance measurement. The first paper reviews the individual-level issues to consider when evaluating PRO instruments for use in patient outcomes. This paper, the second in the series, is intended to help inform next steps on the path to developing PRO-based performance measures (PRO-PM) that can be used at the provider level to assess organizational-level quality of care. Included will be a review of the key methodological issues and a discussion of best practices for meeting NQF criteria for these PRO-PM.

The NQF has undertaken extensive work in the area of outcomes over the last few years. The report on the *National Voluntary Consensus Standards for Patient Outcomes: A Consensus Report*,[1] defined outcomes as being important, because they "reflect the reason that an individual seeks healthcare services." The patient's voice in these outcome measures, however, has largely been missing. There are even fewer PRO-based performance measures at the organizational level, even though patients are often the best able to report on the experiences and results of their individual care.[1]

This paper, by focusing on measures of provider performance, examines some key methodological issues and concerns, several of which also apply to individual-level item or instrument development (i.e., issues of reliability and validity). The instrument must be a valid measure of the desired concept and be reliable when used repeatedly. However, when aggregating the individual instrument responses across all cases for a provider, there are additional considerations of validity and reliability associated with how the measure is constructed and interpreted. These include considerations of the appropriate population, exclusion criteria, the calculation of the performance score, and the use of risk adjustment to adjust for population case-mix differences.

PRO-PM are even more complex as they rely on the patient's subjective assessment of quality which may be affected by their expectations and other perceptions that may differ from the clinician's expectations. While the patient voice is important, particularly as it provides the best measure of whether the patient seeking care perceived the treatment to be effective, it may be less objective in terms of achievable goals. Hence, PRO-PMs may differ in their usefulness for a quality improvement program compared to a regulatory mandate designed to ensure at least minimal levels of quality.

Given the multi-disciplinary nature of this work, we will first define a few key terms used in this paper as defined in prior NQF documents:

Patient-reported outcome (PRO): Any report of the status of a patient's health condition, health behavior, or experience with healthcare that comes directly *from the patient, without interpretation* of the patient's response by a clinician.

PRO patient-level measure/instrument (PROM): *Tools* to assess health condition at the individual level (e.g., health status and status of physical, mental, and social functioning, health behavior, or experience with healthcare).

PRO-based performance measure (PRO-PM): An organizational performance measure based on patient-reported outcome data aggregated for an accountable healthcare entity (e.g., percentage of patients in an organization whose depression score as measured by the PHQ-9 improved).

Performance measure: *Numeric quantification* of healthcare quality for a designated accountable healthcare entity, such as a hospital, health plan, nursing home, clinician, etc.

Provider: A clinician, facility, or organization.

Reliability refers to the *repeatability or precision of measurement*. Reliability of data elements refers to repeatability and reproducibility of the data elements for the same population in the same time period. Reliability of the measure score refers to the proportion of variation in the performance scores due to systematic differences across the measured entities (or signal) in relation to random error (or noise).

Validity refers to the *correctness of measurement*. Validity of data elements refers to the correctness of the data elements as compared to an authoritative source. Validity of the measure score refers to the correctness of conclusions about the quality of entities that can be made based on the measure scores (i.e., a better score on a quality measure reflects higher quality).

Key Components of a Performance Measure

An organizational performance measure is a numeric quantification of healthcare quality for a designated accountable healthcare entity, such as a hospital, health plan, nursing home or clinician. When evaluating a performance measure, it is useful to consider its key components: 1) the item or instrument that measures the health concept of interest; 2) the performance score; 3) the target population and exclusion criteria; and 4) the risk adjustment methodology.

For a PRO-based performance measure, the selection of the items or the instrument that measures the health concept at the individual level-- the PROM -- is central to the PRO-PM's or performance measure's scientific acceptability. In the paper "Methodological Issues in the Selection, Administration and Use of Patient-reported Outcomes in Performance Measurement in Health Care Settings," Cella et al.[2] identified the following eight important characteristics to consider when evaluating and selecting PROs for use in performance measures: 1) conceptual and measurement model, 2) reliability, 3) validity, 4) interpretability of the scores, 5) burden, 6) alternative modes and methods of administration, 7) cultural and language adaptations, and 8) electronic health records (electronic data capture).

While all of these features are important in developing PRO items as they may not all be appropriate for PRO-PM. For example, pain and depression symptoms are important health issues to measure as PRO, but a method to collect these data accurately from a proxy through observation of the patient has been challenging to develop. Hence, building a PRO-PM across all relevant patients in an organization maybe more complicated than assessing outcomes at the individual-patient level. Similarly, patient experience is another important PRO concept, but patient perceptions of the provider should not be incorporated into a patient's EHR. Each of the eight PROM characteristics should be considered in the evaluation of a PRO-based performance measure, but a PRO-PM that has evidence of many, but not all characteristics, should not be rejected from consideration.

On the other hand, five of the eight characteristics address critical issues which should be required evidence in a PRO-PM: conceptual and measurement model, reliability and validity for the target population, interpretability of scores, and burden.[2] These 5 characteristics map to NQF's evaluation criteria of importance (conceptual and measurement model), reliability, validity, feasibility (burden) and usability (interpretability of scores). When evidence supporting the 3 additional criteria, alternative data collection methods (sources, modes and methods of administration), cultural and language adaptations, and inclusion of the PROM data in EHRs (if appropriate),[2] are available, the PRO-based performance measure may be more robust.

A second component of the performance measure is the performance score, which is a provider-level value that is meant to distinguish between good and poor quality, and is derived from the individual-level PROM data. The PROM data used in performance measures may be the score from an individual item (e.g., pain score) or a scale score derived from an instrument or set of items (e.g., PHQ-9 score). The PROM data are aggregated to the provider level and the performance score can take the form of a count, a percent, a mean, or a ratio value. Descriptions and examples of these performance score options are provided below as part of the performance measure validity considerations.

A third key component of a performance measure is the target population of patients to be included in the performance measure and the exclusion criteria. The target population selected for a PRO-based performance measure should be based on literature documenting the use of the PROM in the target population. Some PROMs are classified as generic and would apply to a wide range of individuals, while other PROMs are condition-specific and will be appropriate for use with a limited target population.[3]

Risk adjustment is a fourth key component of a performance measure. Most, but not all, outcome measures use risk adjustment to account for case-mix differences across providers or across time so that observed differences can be attributed to the provider's care and not population differences.[4, 5]

All these components -- the selected PROM, the performance score derived from the PROM data, the target population and exclusion criteria and the risk adjustment methodology -- contribute to the PRO-PM or the performance measure's ability to make a valid estimate of the quality of care furnished by a provider. In the next two sections, we build on the NQF's Measure Testing Task Force Report[6] and describe issues related to the reliability and validity of the performance measures and, where appropriate, identify unique issues that are specific to performance measures based on PROMs rather than clinician assessment of outcomes or process descriptions.

Reliability

Reliability is an important measurement concept as it refers to the repeatability or precision of measurement. *Reliability of data elements* refers to repeatability and reproducibility of the data elements for the same population in the same time period. *Reliability of the performance measure* at the provider-level refers to the proportion of variation in the performance measure due to systematic differences across the measured entities (or signal) in relation to random error (or noise).[6] It is important to note that data element reliability does not guarantee reliability of the provider-level performance measure (clinician, facility, or organizational-level). Also, reliability is a necessary, though not sufficient, pre-condition for validity, whether looking at individual-level PROM scores or at PRO-PM.. Lack of reliability in performance measures can result in misclassification of providers in quality rankings, which could have adverse impacts on public reporting, perceptions of provider quality, and pay for performance.[7, 8] In this section we expand on the definition of reliability given above, and describe methods for evaluating the reliability of provider performance measures, and strategies for improving reliability when test results show measure reliability may be a concern.

Patient-Level Reliability and Provider-Level Reliability

At the patient-level, reliability can be described as the ratio of the variability among patients' PROM scores to the total variability of the PROM scores, which can also be stated as reliability = subject variability/(subject variability + measurement error).[9] The formula at the performance measures level is conceptually the same. Reliability can be represented by the following formula: reliability = signal/(signal + noise), which is described by Adams as "the squared correlation between a measurement and the true value of the measure."[7] Reliability has a range of 0 to 1, with 0 indicating that all variability in a measure can be accounted for as measurement error, and 1 indicating that all variability is due to actual differences among providers.[7, 10]

Provider-level reliability is determined by the: 1) magnitude of true differences among providers; 2) within provider-variation; and 3) the size of the provider sample, or denominator.[7] Reliability is not an intrinsic characteristic of a measure, but dependent on the characteristics of the set of providers and patients included in the measure specifications. Reliability is also not static; if all providers improve their performance, resulting in reduction in variance among provider performance measure scores, reliability will decrease.[7] Also, estimates for smaller providers are more vulnerable to random error than are estimates for larger providers.

Reliability Testing

There are multiple methods for testing reliability of the provider-level performance measure (e.g., signal to noise). Two-level hierarchical models can be used to estimate signal and noise. A tutorial for this approach has been published by RAND.[7] These models control for random error at the patient-level and at the larger, organization level. For binary patient outcome measures, signal can be estimated with a hierarchical logistic regression model, obtaining the value for 'signal' from the variance of the provider random effect. Noise is calculated based on the standard error of a proportion.[7, 11] Thresholds cited in the literature for reliability for performance comparisons among groups is 0.70, and for individuals is 0.90.[7] Other types of hierarchical models can be used for other types of performance measures (e.g., hierarchical linear model for composites or continuous measures). Regardless of model type, this method of using hierarchical modeling results in a reliability estimate for every provider. If a large proportion of providers fall below a threshold of 0.70, reliability may be a concern for the performance measure.

There are also other potential strategies for quantifying the random error around a performance measure. This includes reporting provider-level performance measures with an estimate of uncertainty, such as a confidence interval for each provider's performance measures. This allows for examining random error relative to the differences in provider-level scores, by examining the overlap of confidence intervals.[12] The specification of the performance measure determines the appropriate approach for calculating the confidence interval.[12] Another systematic strategy for estimating reliability is to calculate what Zaslavsky[12] calls the interunit (here we would call it inter-provider) reliability or the proportion of the variance in a measure that is due to true differences in provider performance. This can be calculated using the value of F derived from an F-test. Interunit reliability can be specified as 1-(1/F) and can be interpreted similarly to the reliability score described above. The F value can be obtained using statistical procedures such as SAS PROC GLM.[12] Note however, that, unlike the method using hierarchical models described in the prior paragraph, this interunit reliability calculation does not result in a reliability estimate for each provider. This summary metric can be calculated when the standard errors of estimates for providers are similar.[12]

The intra-class correlation coefficient has also been used to calculate measure reliability.[13] Hofer describes calculating reliability based on the mean of patient values within provider and the inter-class correlation coefficient using the Spearman-Brown prophecy formula.[14] The interpretation is similar to the other reliability indexes described above, as the provider (or subject) variance over the total variance, which includes the provider variance and variance attributable to measurement or random error. For example, one would interpret that 30 percent of variation in a measure is due to chance, given a calculated reliability of 0.70.[14] Other methods for evaluating performance measure reliability include generalizability analysis based on generalizability theory on sources of variation, which uses a factorial analysis of variance (ANOVA).[15, 16] This method allows for the calculation of components of variance. Though potential sources of error need to be measured, the measure also includes an interaction component, which takes into account variation between sets of patients and providers, and error variance that is attributable to the design of a measure, and residual variance which is error from sources that are unmeasured.[15, 16] Results can be used to quantify the proportion of variation in a measure that are attributable to error, and also potentially provide targets for improving reliability by decomposing error into its potential sources.

The last method for evaluating reliability that we describe is Monte Carlo simulation. Monte Carlo estimation also allows for the evaluation of the performance measure validity when the true value of a provider's performance measure is not known. [17] For example, in a study that examined the reliability of four Bayesian measures of hospital mortality, the reliability of the measures was estimated by examining the correlation between pairs of the measures based on 100 Monte Carlo simulations.[17] Reliability results showed similar reliability to the most frequent measures of mortality evaluated also using Monte Carlo simulation methods.

Sample Size and Reliability

Another important question to consider in performance measure design is the sample size needed to provide a reliable estimate and the likelihood of providers achieving an adequate sample size for a specific level of analysis (e.g., clinician or hospital)? Consideration should be given to the minimum sample sizes needed to reliably calculate provider-level (individual provider or organization) scores on PRO-based performance measures and understanding how many providers would fall below this minimum sample size. Because reliability is not dependent on sample size alone, but, as described above, also dependent on the real differences between providers, and within provider variation, sample size cutoffs commonly used to suppress provider results, most vulnerable to random error, may be insufficient.[7] Signal-to-noise calculation of reliability can allow some estimation of a minimum sample size necessary to meet reliability thresholds listed above. Examining the relationship between provider reliability estimates and their measure denominator sizes can be useful for identifying the threshold size where reliability estimates are 0.70 or greater.[7] [11]

If reliability estimates show that a measure has poor or marginal reliability (i.e., the reliability indexes described above are below 0.70), or if large proportions of providers have reliability estimates below this threshold, it may be necessary to consider methods for improving the reliability of the performance measure. Potential strategies for improving provider-level measure reliability can include designing composites, which combine provider scores on more than one performance measure, thereby increasing data points being used and increasing the performance measure stability.[13] An additional strategy is to increase provider-level sample or measure denominator size by increasing the performance measure time period; adding time periods to the denominator increases the number of cases included in the measure denominator. A potential drawback to this strategy is that the measure may be insensitive to changes in quality over time.

Reliability Adjustment

Another strategy for reducing random error in a performance measure is to use shrinkage estimates towards the mean value for providers. For smaller providers, which are more vulnerable to random error, the shrinkage towards the mean is greater. Estimating performance measures using hierarchical modeling with empirical Bayes shrinkage estimators is one strategy for accomplishing this.[10, 12, 18] One potential concern about this strategy, which should be noted, is that the shrinkage estimate may mask a poor provider's performance. If the true performance of a provider is quite poor compared to the mean, the shrinkage estimate will mask that difference by pulling the provider performance measure score towards the mean. Additionally, smaller providers that have values that indicate higher quality than the mean, will be pulled closer towards the mean.[10] An alternate strategy to potentially handle these criticisms is to shrink estimated provider performance measures towards the mean value expected for a provider of that size.[10] However, given the evidence available connecting higher volume to quality, there is debate in the literature about using provider size for reliability adjustment, if size is being measured as volume of patients. Volume is potentially endogenous because prior low quality may be the cause of the current low volume, though the debate over the use of volume alone is not settled.[10] An additional strategy available to address this criticism is shrinking the provider performance measure towards the mean value expected for a provider of that size and other attributes .[19] Alternatively, if size can be measured using a different metric, such as numbers of beds, the potential endogeneity can be reduced.[19] Examples of this method of using size and other provider attributes for reliability adjustments applied to performance measures are available.[19] Other limitations to the strategy of shrinkage towards the mean are cited in the literature as well, for example, examining hospital mortality performance showed that this strategy did better at identifying the "best" hospitals than it did identifying the "worst."[10]

These approaches to assess the reliability of performance measures apply to PRO-based measures in the same manner that they apply to other types of performance measures. The best approach, as noted above, depends on the specific analytic conditions, and will differ depending on the attributes discussed above.

Validity

A second issue that needs to be addressed within the scope of scientific acceptability of the performance measure is validity. As noted in the NQF Measure Testing Task Force Report,[6] validity refers to the correctness of measurement. *Validity of* data elements refers to the correctness of the data elements as compared to a "gold standard." *Validity of the performance score* refers to the correctness of conclusions about the quality of the provider that can be made based on the performance scores (i.e., a better score on a quality measure reflects higher quality). Therefore, item- or instrument-level validity of a PROM is necessary, although not sufficient, for its use as a performance measure, or PRO-PM. Use of an instrument that is not reliable and valid would mean that the performance measure would not measure quality consistently or accurately. It is also important to note that if an item, instrument, or performance measure is not reliable, it cannot be valid.[9] The methods used to calculate the performance score, the target population (denominator), and the risk adjustment procedures will determine whether a valid and reliable instrument can be a used to create a valid and reliable PRO-based performance measure.

PRO-Specific Validity Issues

For each PRO concept, unique features will need to be considered as the PROM data are used to develop a PRO-PM. For example, for PROMs addressing depression symptoms, a unique feature is the need to ask respondents to reflect on a 14-day look-back period and the relatively long time needed to observe benefits from a treatment plan. Therefore, any performance measure that addresses the effectiveness of treatment for depression must consider a reasonable treatment effectiveness time frame *and* the PROM time frame of 14 days. The performance measure called "Depression Remission within 6 Months" does recognize these time frames, and collects follow-up data at 6 months.

Another important issue related to PRO-PMs like depression is that instruments that measure symptoms, such as depression symptoms are screening tools and are not equivalent to a clinical diagnosis of depression. Therefore, treatment and performance needs to measure symptoms within those with a clinical diagnosis, rather than a screening measure. The performance measure "Depression Remission within 6 Months" handles this issue through its denominator inclusion criteria which requires a patient to have a clinical diagnosis of major depression or dysthymia (based on IDC-9 codes) *and* a PHQ-9 score that is higher than 9. This ensures that only those patients with the clinical diagnosis who should be treated are included in the performance measure focused on the effectiveness of treatment.

There are also unique features to consider regarding the PRO concept of pain. Pain management may be a primary treatment goal for certain populations, such as patients with low back pain, and a PRO-based performance measure targeted to this population may be focused on reduction of pain between the start of treatment and the end of treatment. Pain is also a general symptom that is often monitored across all patients, and a performance measure based on pain can be applied to an entire population of a provider. For example, the performance measure "Percent of patients with moderate to severe pain (long stay)" is targeted for nursing home residents and focuses on the percent of patients having a high level of pain, which is defined in this measure as constant or frequent pain and at least one episode of moderate to severe pain or very severe/horrible pain of any frequency. The pain measurement field is moving from just using a numerical rating scale that does not take into account individual pain tolerance to measures of the pain's effect on sleep or other activities

(See recent work by the Centers for Medicare and Medicaid Services on measuring pain in the CARE item set (http://www.pacdemo.rti.org/).

Within the domain of health behaviors, which includes risk behaviors that are potentially detrimental to health, such as smoking and excessive alcohol intake, the initial prevalence of the behavior can be expected to vary by geographic region, and thus by provider (http://www.cdc.gov/nchs/data/series/sr_10/sr10_252.pdf). Provider-level comparisons of reductions in smoking would be affected by the initial prevalence of smoking, and the number of patients in the target population could vary a great deal across providers; many providers may have very small number of smokers. If only a small number of patients within a provider meet the criteria for inclusion in a performance measure, the performance measure score may have problems with reliability and therefore validity.

For the concept of health-related quality of life (HRQL), a multi-dimensional construct that covers physical, social, and emotional well-being,[20] a challenge will be to include the HRQL instrument measure within a performance measure in a way that reflects the quality of services furnished by the provider. Evidence documenting healthcare interventions that lead to improved HRQL in the target population will be needed to demonstrate validity. Performance measures based on HRQL may best be targeted to special populations, for example individuals with a spinal cord injury.[21]

Selecting Aggregated Provider-Level Measures: Change Scores versus Threshold Values

As noted in the description of the key components of a performance measure, there are several options for aggregating patient-level PROM data into a provider-level PRO-PM. A key question considers which method of aggregating data at the provider level will be most discriminating in terms of noting differences in quality? The approach to creating a discriminating performance measure will vary depending on the PROM. The expected outcome measure for PROMs will be either a desirable or undesirable outcome or health status.[22] There are several options for reporting the provider-level outcome or health status including: 1) a change in status (e.g., decrease in pain between start of care and end of care, increase in functional status between start and end of care) if the instrument is sensitive to change (i.e., responsive)[22, 23] or 2) a threshold achieved (e.g., percent of patients with moderate to severe pain).[23] While the change in health status, calculated as the difference between follow-up score and baseline score, may initially seem like the best choice for measuring the effectiveness of care, such as reduction in pain or improvement in functional status, there are several methodological limitations with calculating change scores.[9] One problem with calculating a mean change score at the provider level is that individuals' change scores can vary in the magnitude and the direction of the change, and these individual differences could be masked as positive and negative changes cancel each other out. Change scores are also subject to measurement error from the baseline scores and error from the follow-up scores, so change scores tend to have lower reliability than the baseline and follow-up scores. Another challenge to measuring change relates to item or instrument floor effects for patients who start at the low end of a scale and ceiling effects for patients who start at the high end of the scale. An individual may have an improvement in health, but the instrument cannot detect the change because the instrument is not sensitive enough at the low end or high end of the scale. Finally, for many health status measures, the clinical meaning of a change score is unknown, and the change score is hard to interpret. Ideally, health status measures would have clinically-meaningful thresholds that could define "stages" or "complexity" levels,[24-29] and moving from one clinically meaningful stage to another stage would have meaning. The PRO-based performance measure "Change in basic mobility as measured by the AM-PAC" uses a change score to document improvement in functional status, while the PRO-based performance measure "Percent of patients with moderate to severe pain" uses a threshold value. The performance measure "Depression Remission within 6 Months" uses a threshold value, but reflects a change from a baseline PHQ-9 score indicating possible depression, to a follow-up up score indicating no depression (i.e., remission). The best approach really depends on the measurement goal.

Defining the Performance Score Using Aggregated Data

The performance score is calculated at the provider-level and may be a mean, percent, or ratio value. It is meant to distinguish between good and poor quality, and is derived from the individual-level PROM data. For a PRO-PM score or measure that is continuous, a summary statistic of central tendency, such as a mean, can be calculated as the performance score. Calculating a mean takes advantage of all the detailed amount of data available for analysis. However, if individual-level data within the provider tend not to be normally distributed, a mean may not be the best estimate of central tendency as it will mask variation within the provider. A mean value may also be misleading in situations where the population is heterogeneous because it does not represent the diversity of the patient population. [30]

An alternative to reporting a mean is to calculate a percentage value based on the number of patients who achieve or exceed a specified benchmark. The benchmark may be defined in various ways, including: 1) a national expected value (either threshold or change) based on outcomes of similar patients; 2) a fixed amount of change defined based on a PROM-specific clinically important difference or PROM-specific minimal detectible change; or 3) a threshold value that is associated with a longer-term outcome (e.g., balance score associated with a reduced risk of falls). For PROMs that have established clinically meaningful thresholds (i.e., cut points), the performance score should incorporate these thresholds. For PROMs that do not have established clinically meaningful groups or thresholds, establishing validity of the performance score will be more challenging. For the PHQ-9, which measures depression symptoms, a score ≥ 10 has been found to be clinically important, with a sensitivity of 88% and a specificity of 88% for a clinical diagnosis of major depression.[31] Sensitivity refers to how often the test will be positive (true positive rate), if a person has a condtion. Specificity refers to how often the test will be negative (true negative rate) if a person does not have the condition. For the performance measure "Depression Remission within 6 Months," the depression measure does not use a mean or median change in the PHQ-9 score, but rather classifies scores into clinically meaningful groups (< 5 = not depressed and > 9 = depressive symptoms), and the patient is considered to have made an improvement if he or she moves from the depressive symptom category (> 9) to the not depressed category (< 5). The performance measure "Change in basic mobility as measured by the AM-PAC" includes admission and discharge mobility measures and the performance score is the percent of patients with a change/improvement. Change for this performance measure is defined as a difference of one or more minimal detectable change(s). A minimal detectable change refers to the minimal amount of change that is not likely to be due to measurement error, and thus represents a true change.

A third option for the performance score is a ratio value, which is a score that may have a value of zero or greater that is derived by dividing a count of one type of data by a count of another type of data (e.g., the number of patients reporting pain score of 7 or higher divided by the number of inpatient days). A ratio may be preferred as the performance score when the amount of time (e.g., number of days) that a patient is at risk for the outcome is important.

Again, each of these approaches have been used in different measurement efforts. Selecting the right one depends on the type of performance being examined. Consensus on which method is best really depends on the type of question being answered and whether the proposed approach provides the most robust measure possible given the constraints of the measurement design.

Methods of Validity Testing

A second question related to validity of a PRO-based performance measure considerswhat methods f would support the demonstration of validity of the provider-level performance score for making conclusions about quality of care? As with other performance measures, validity testing is very important, but strong evidence at the performance measure level can be challenging to obtain. Validity testing often begins with face validity, which refers to the clinical credibility of the measure based on expert review. Face validity can be tested using a systematic process such as a modified Delphi survey,[6] a formal consensus process, use of the UCLA/RAND Appropriateness Method[32] or the American College of Cardiology and

American Heart Association Methodology for the Selection and Creation of Performance Measures.[33] Given that PROMs represent the patient's perspective, face validity of PRO-based performance measure could also be tested with "patient experts" using qualitative research methods, such as focus groups, semi-structured interviews, and cognitive interviews. If patient experts are used, it will be critical to describe and frame the concept of healthcare quality in a way that individuals understand this complex issue. Hibbard[34] provides a foundation for this framing. Although face validity is generally not considered to be strong evidence of validity, it is important to have input from experts outside the research or measure development team review the measure specifications.

Validity may also be tested based on criterion validity, which refers to the extent that the measure agrees with a "gold standard." This may include another measure of the same construct collected at the same time (concurrent validity) or correlation with another measure, such as a longer-term outcome (predictive validity).[6] For a PRO-based performance measure, comparisons of a performance score based on clinician observation that taps into the same construct (e.g., functional status) may be one way to demonstrate concurrent validity. Finally, construct validity of a PRO-based measure may be established. Construct validity refers to how the measure performs based on theory.[6] Construct validity could potentially be tested by identifying providers who implemented quality improvement (QI) initiatives focused on a PRO construct and comparing the providers' performance scores before and after the QI program. If provider performance scores do change as expected, and if performance scores from "control" providers that did not have a QI program focused on this construct do not change, this could strengthen findings.

Another validity issue in PRO-PM must consider how patient preferences are taken into account. For example, a patient may report a high level of pain but prefer not to have pain medication or alternative pain management treatments. Using a change in pain level or a threshold may not recognize patient preferences related to pain management. An example of a PROM that addresses patient preferences related to pain can be found in the Family Evaluation of Hospice and Palliative care survey. The family member or significant other is asked: How much medicine did the patient receive for his/her pain? The response options are: 1) Less than was wanted; 2) Just the right amount; and 3) More than the patient wanted. In this survey, the respondent is asked a question about the treatment within the context of the patient's preference. A performance measure may also account for patient preferences by excluding these patients from the denominator. However, as noted in the CSAC Guidance on Quality Measure Construction,[35] the effect of exclusions for patients preference should be transparent, because exclusions for patient preferences (e.g., refusal) may be related to quality problems.

Risk Adjustment

Risk adjustment is an important factor for developing outcome measures, including PRO-PM. The clinical outcomes of care, both the desirable and undesirable outcomes, are often a result of patients' personal and clinical factors, as well as the quality of healthcare services. Differences in patient case-mix exist across providers because patients are not randomly assigned to their healthcare providers. Therefore, when patient outcomes of care are compared across time or across providers, these outcomes often need to be adjusted to control for patient-level factors so that the effect of the providers' care can be isolated. The purpose of risk adjustment is to allow for a "fair" comparison of health outcomes, so that observed differences can be attributed to the provider interventions and not population differences.[4, 5] Domains to consider for risk adjustment of PRO-based performance measures and different approaches to risk adjustment are described below.

Selecting Factors for Risk Adjustment

Patient factors selected for risk adjustment of a PRO-based performance measure should be based on evidence that the factor affects the outcome independent of the intervention. Evidence would include peer-reviewed research literature as well as clinical expert opinion. Informed patients could provide very valuable insights into potential covariates. Covariates will be different for different PRO concepts. For example, factors associated with higher risk of pain might be severity of arthritis or time since surgery while factors associated with functional status might include primary diagnosis, age, baseline functional status and comorbidities.

Several factors often used in risk adjustment can be generally categorized into patient demographic factors and patient clinical factors that are present at the start of care. Demographic characteristics, such as age, are often included in risk models. Clinical factors such as diagnosis, severity of illness, comorbidities, and baseline scores that affect outcomes are also often included in risk adjustment models. The relationship between the baseline and follow-up (threshold) scores can affect analysis results. If the correlation between baseline scores and follow-up scores are high, then a change score is more likely to be significantly different than a follow-up. If the correlation between baseline and follow-up scores is low, then the follow-up score is more likely to be significant. [9] Psychosocial factors, such as adherence, motivation, understanding, engagement, and readiness to change, have been suggested as potential covariates for PRO-based performance measures. Psychosocial data has not typically been available for patients, and so use of these factors as covariates would likely require additional data collection. In addition, the inclusion of psychological factors in models may be controversial. For example, physical therapists at a provider may be very skilled at engaging and motivating patients to be physically active, and their patients may report superior functional status outcomes compared to other providers. If patient motivation was a covariate in a risk adjustment model, motivated patients would be expected to have superior outcomes, and the therapists' ability to motivate patients would be masked.

There is controversy surrounding risk adjustment of patient factors such as race, ethnicity, socioeconomic status (SES), and limited English proficiency, which have been associated with poorer outcomes and also with disparities in care. These factors are not typically included in risk models for performance measures. Including factors associated with disparities such as race, ethnicity, SES or limited English proficiency in risk adjustment models could mask quality problems due to disparities[35] and would suggest that differences in outcomes based on these patient factors are acceptable, and do not need to be eliminated. NQF's guidance for measure evaluation indicates: "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, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences." Many measures have not adjusted for these factors or stratified data by these factors. Several stakeholder groups have expressed strong concerns about not adjusting for these factors, because it may lead some providers to avoid admitting these patients ("cherry-picking"), and thus limiting access to care for low-income and minority patients. This may also lead to the concentration of low-income and minority patients receiving care from providers that have that may have fewer resources. In their paper focused on Healthcare Disparities Measurement, Weissman et al., suggest a combination approach to the issue may be needed. They offered 2 recommendations:

1) Stratification by race/ethnicity and primary language should be performed when there are sufficient data to do so. Risk adjustment may be appropriate when performance is highly dependent on community factors beyond a provider's control.

2) Performance reports stratified by race/ethnicity should not be risk adjusted by SES or other contributory factors, and instead could be stratified by SES if the data permit.

Risk Adjustment for Alternative Data Collection Sources, Methods and Modes

A unique feature of PRO-based performance measures is the process of data collection, which may need to be flexible in order to accommodate patients' diverse needs. Flexible data collection methods may mean that data are not equivalent, and risk adjustment methods may be appropriate to adjust for systematic differences tied to data collection methods. Data collection may vary in 3 key ways: 1) the source, 2) the mode, and 3) the method. The source of the data for a PROM will most often be the patient, but in some cases a proxy may be needed to report on behalf of the patient. The mode of administration, either self-administration or interviewer administration, may also vary. Patients may not complete a survey without someone asking the questions and recording their responses. A third factor that may vary is the method of administration, which could include paper and pencil, telephone, or computer. If the source, mode or method of data collection varies across providers or across time, the provider-level performance measure data may not be comparable. If research comparing the alternative source, mode or method shows that the PROM data are equivalent, then data could be pooled regardless of data collection source, mode, or method. If, however, data are not equivalent, research may identify an adjustment factor for alternative data collection procedures, and the adjustment factor could be part of the risk adjustment model.[36, 37] If research has not found the scores derived from varied data collection methods to be equivalent or an adjustment factor has not been identified, because differences are not predictable, data could not be pooled and the data for these patients would be missing for the performance measure. For the PHQ-9 instrument, a clinician observation version has been developed, and initial feasibility testing has been conducted; however, equivalency of the self-report and clinician observation measures has not been established.[38, 39] For the performance measures "Depression Remission within 6 Months," and "Percent of Residents with Moderate to Severe Pain (short stay)" patients who cannot self- report are excluded from the performance measures.

Other data collection issues that may require adjustment include child health PROs, where either the patient or the parent may be the expected source,[40-44] and different language versions of a PRO instrument that may result in responses that are systematically different. Again, research may support pooling these data, either with or without adjustment. If evidence does not support pooling the data, because the alternative source or language version is not comparable, use of the alternative would be considered missing data. To increase the likelihood that different language versions of a PRO instrument do lead to equivalent patient-level scores, principles of best practices for translation and cultural adaptation should be followed. Wild et al. provide suggested best steps for translation and cultural adaptation of PROM instruments.[45, 46] It is worth noting that there are now more than 75 translations of the PHQ-9 available on the internet (http://www.phqscreeners.com/overview.aspx).

Risk Adjustment Methodology

There are several approaches to risk adjusting outcomes data. One option is to identify high- and low-risk groups and report the data stratified by these risk groups (e.g., strata). Outcomes for patients across providers could be compared within the same strata. This approach is appropriate when adjustment for one key factor is needed, and the factor is either dichotomous or has clinically-meaningful cut points (and could be made dichotomous) and when the number of patients is large enough to split them across 2 or more groups. A second risk adjustment approach uses regression modeling with demographic, medical, and data collection (if appropriate) factors included in the model as covariates. With this approach, multiple factors, including continuous and dichotomous factors, can be controlled for, and facility-specific predicted and expected values can be calculated in order to compare risk-adjusted data across providers.

A third risk adjustment approach would involve identifying risk groups (i.e., strata), *and* using regression models within each strata. These data are then aggregated into a single estimate based on the national distribution of patients by strata. This combined approach would be needed if the effect of key covariates on the outcome varied by strata (risk) group. For example, if the effect of baseline functional status on discharge functional status varied by primary diagnosis, then data should be stratified by diagnosis and regression models for each diagnosis group would be used. The regression results would be aggregated into a summary score based on weighting of the diagnosis groups (strata) using a national distribution or other standard. For condition-specific measures, when the target population has a common medical diagnosis, regression modeling may be adequate to adjust for several covariates such conditions severity, age, and comorbid conditions. When the target population for the performance measure is heterogeneous, then the combined approach of strata and regression modeling may be the best option.

A significant area of controversy is the choice of the type of regression model used in the risk adjustment process. Concerns about clustering and small sample sizes within providers have led some measure developers to use hierarchical generalized linear models (HGLMs), rather than fixed-effects regression models. The HGLM approach has been criticized, because it decreases the variation in the performance score, particularly for small hospitals. In the paper "Statistical Issues in Assessing Hospital Performance," commissioned by the Committee of Presidents of Statistical Societies, Ash and colleagues critically reviewed this issue and indicated that HGLMs are appropriate for use given the structure of the data and the purpose of the analyses.

Although most performance measures that are outcome measures need to be risk adjusted in order to make fair comparisons across providers or across time, there are exceptions. For example, if an undesirable outcome should not occur, regardless of patient's demographic or clinical factors, then risk adjustment may not be necessary. A PRO-based performance measure that is not risk adjusted is the measure "Percent of residents who self-report moderate to severe pain (short stay)." For this performance measure, the expectation is that no resident should experience severe pain or moderate pain frequently or almost constantly, therefore, the percent of residents who have moderate to severe pain is reported without adjusting for patient or clinical factors.

Threats to Validity

There are many potential threats to validity for performance measures and PRO-based performance measures in particular. Threats to the validity of a performance measure can be classified into three broad categories, including invalid item or instrument for the target population, missing data due to non-response or other reasons and inadequate case-mix adjustment.

Threats to Validity: Item and Instrument Validity

Factors affecting the PROM item/instrument's reliability or validity can threaten the validity of a performance measure based on that PRO. For example, patient responses may shift over time, but not because of true change.[9] Patients may not give accurate responses because of social desirability concerns.[9] Patients may also have a tendency to give positive or negative ratings for patient experience measures,[47] and an uneven distribution of these patients may affect providers' performance measure estimates. For PROMs that are interviewer administered, inter-interviewer variability is also a potential concern. The PRO-based performance measure "Percent of residents with moderate to severe pain (short stay)" relies on data collected by interview. The pain data are collected using the Minimum Data Set 3.0, a patient assessment instrument that is required by the Centers for Medicare and Medicaid Services. A script for asking the patient about pain is included on the MDS 3.0 form, and this may support inter-interviewer reliability. The PRO-based performance measure called "Depression Remission within 6 Months" uses clinical diagnostic data to address the validity of the PHQ-9 score, specifically, the baseline depression symptom score. For this performance measure, the target population is patients who are depressed. The patients are classified as depressed based on the initial score from the PROM PHQ-9 instrument. A patient may have a PHQ-9 score of greater than 9 during the initial assessment, but is not clinically depressed. The performance measure specifications require a clinical diagnosis of depression, in addition to the PROM depression score, in order for the patient to be included in the denominator. This means that patients who have a PROM score suggesting depression, but are not clinically diagnosed as depressed, are not included in the denominator and thus, not included in the calculation of remission at 6 months.

If computer-adaptive testing is used to collect data, the PRO instrument should have been tested for differential item functioning within subgroups of the target population. Otherwise, individuals may be assigned a value that does not reflect their true health status. For example, motor functional status might best be separated into the constructs of self-care and mobility rather than one single construct of motor function (which combines self-care and mobility) when the population is heterogeneous. This will allow the functional outcome value to vary depending on their ability within each subscale. Given that individuals recovering from a hip replacement and those recovering from a central cord spinal cord injury will have different patterns of motor ability (differential item functioning) that can be differentiated using the two subscales of self-care and mobility skills. Use of different PRO instruments to measure the same construct, such as depression symptoms, could be used, but research demonstrating the agreement with assignment to clinically important groups (e.g., depressed, not depressed) should be high. If the research examining the equating process shows the assignment into clinically meaningful groups is not well aligned, this may introduce systematic errors based on the instruments selected. The performance measure "Percent of residents with moderate to severe pain (short stay)" allows for pain data to be collected based on the numeric rating scale (0 to 10 scale) or the pain verbal descriptor scale (mild, moderate, severe, very severe/horrible). The performance measure equates thresholds of pain across the 2 items.

Threat to Validity: Missing Data

A second threat to validity occurs when data are missing but not missing at random. As noted in the CSAC Guidance on Quality Measure Construction,[35] missing data may be indicative of a quality problem itself, therefore, excluding those cases may present an inaccurate representation of quality. For PROM data, a key issue is response rates of patient surveys. During the testing of a PRO-based performance measure, response rates would be important to monitor and report. A survey with a low response rate during testing (somewhat ideal circumstances) would likely have lower response rates in clinical practice. If response rates are low, and the individuals who do not respond are different than the individuals who do respond, non-response error is a concern.[48] An additional concern about response rates is that they are often not calculated correctly and are sometimes misrepresented.[3] Standard definitions with calculations have been developed by the American Association for Public Opinion Research (AAPOR) and one or more of these definitions could be adopted for PRO-based performance measure testing. The AAPOR Council has indicated that no single number or measure reflect the quality of a survey and provides the following four definitions and formulas for calculating response rates, cooperation rates, refusal rates, and contact rates:

Response rates - The number of complete interviews with reporting units divided by the number of eligible reporting units in the sample. The report provides six definitions of response rates, ranging from the definition that yields the lowest rate to the definition that yields the highest rate, depending on how partial interviews are considered and how cases of unknown eligibility are handled

$$Response\ Rate\ 1 = \frac{I}{(I+\ P) +\ (R + NC + O) +\ (UH + UO)}$$

I = Complete interview

P = Partial interview
R = Refusal and break-off
NC = Non-contact
O = Other
UH = Unknown if household/occupied HU
UO = Unknown, other

Cooperation rates - The proportion of all cases interviewed of all eligible units ever contacted. The report provides four definitions of cooperation rates, ranging from a minimum or lowest rate, to a maximum or highest rate.

$$Cooperation\, rate\ 1 = \frac{I}{(I+P) +\ R + O}$$

Refusal rates - The proportion of all cases in which a housing unit or the respondent refuses to be interviewed, or breaks-off an interview, of all potentially eligible cases. The report provides three definitions of refusal rates, which differ in the way they treat dispositions of cases of unknown eligibility.

$$Refusal\, rate\ 1 = \frac{R}{(I+P) +\ (R + NC + O) + (UH + UO)}$$

Contact rates - The proportion of all cases in which some responsible housing unit member was reached. The report provides three definitions of contact rates.

$$CON\ 1 = \frac{(I+P) +\ R + 0}{(I+P) +\ R + O + NC + (UH + UO)}$$

As previously noted, response rates in a clinical setting would likely be lower than response rates during a research project or testing, and response rates may vary by provider. Given that performance scores may vary at the provider level due to response rates tied to response error, reporting response rates along with performance scores for PRO-based measures may be important.

Although this is not always the case, studies have found that when response rates are low, results are more likely to be biased, either positive or negative.[48-50] Thus, surveys used for PRO-based performance measures should ideally be developed in a way that optimizes response rates. For self-report surveys, simple strategies such as font selection, and the use of check boxes are important.[51] In addition, more recent research[51] has focused on the principle of social exchange, which emphasizes that rewards for responding to surveys should outweigh any perceived costs. For example, Dillman[51] recommends showing positive regard for the respondent by saying thank you and providing a phone number for questions, as well as social validation by communicating that others are participating and that each response is important. The respondents' "costs" can be minimized by keeping the survey short and easy to complete and by minimizing personal information. Trust is another key issue and clearly noting the sponsor of the survey and ensuring confidentiality and privacy of the information provided by the respondent can improve response rates. Ideally, the testing of the survey should have included review by one or more expert panels, consumer input, cognitive interviews, and pilot testing.

Missing data may also be a problem, because patients cannot respond to a survey due to communication limitations, language barriers, physical disabilities, or other reasons. To minimize the amount of missing data, alternative sources (i.e., proxies), and modes and methods of administration (i.e., use of recorders) should be considered. Self administration is often the preferred mode of data collection, because it minimizes interviewer effects on the data and it minimizes burden on clinicians. However, patients may choose not to complete a survey, but would be willing to be interviewed. Overall, comparisons of data collected using self report versus interviewer show high reliability. It is important to note that these studies used trained interviewers who were research staff. In a clinical setting, interviewers would be clinicians rather than research staff members, and clinicians will have varying skills as interviewers and they will be very busy, so the tendency to rush or miss interviews is possible. When data are collected using varying modes or methods, additional PROM-level reliability testing may be appropriate. For example, when data are collected using interviewers, intra-interviewer reliability and inter-interviewer reliability may be appropriate. For the performance measure "Percent of residents with moderate to severe pain (short stay)," data are collected using an interview as part of the mandated Minimum Data Set. This has resulted in relatively low missing data rates.

For patients who cannot respond to a verbal or written survey due to cognitive or communication limitations, a proxy may provide responses on behalf of the patient. In order to use proxy responses within a performance measure, proxy responses would need to be reasonably accurate. Proxies may demonstrate acceptable reliability for PROs such as functional status, where the proxy can observe the patient. However, use of proxy responses are less useful for more subjective PRO concepts, such as pain intensity, nausea, depression symptoms, because proxy data in this area tend to be less reliable.[52] Proxy responses are reasonable to consider for child health measures where parents are proxies and the research has show small differences in child-parent reports. Use of proxies may minimize missing data, but it may introduce error to the performance score, and thus would be a threat to validity.

Threat to Validity: Inadequate Risk Adjustment

Another potential threat to validity might be inadequate risk adjustment methodology. Some providers may have a fairly unique specialty treatment program focused on clinically complex patients (e.g., severe stroke, bariatric patients) and standard risk adjustment methods may not adequately adjust for these uncommon patients' factors. In observation studies, techniques such as propensity score analyses are used to address this problem referred to as selection bias.

Other potential threats to validity include sources of non-random variation, such as seasonal variation, state-level policies, geographic variation in practice patterns, and natural or other disasters (e.g., Hurricane Katrina, earthquakes, etc.).

Conclusion

The area of PRO-based performance measures is relatively young and still evolving. While the patient's voice is often included in experience with care measures, the science of PRO-based performance measures is much less developed. Some work has been done in the areas of pain management and physical health status, but even that work is still evolving to build performance measures that go beyond personal interpretations and instead look at the actual impact on the patient's quality of life. The importance of the underlying science is critical as one moves from measuring outcomes at the individual level to holding organizations accountable for all patient outcomes. While outcome measures can be risk-adjusted, it is difficult to control for the effect of subjective perceptions across all patients in an organization, thus complicating the use of PROs for determining accountability.

This paper discussed the key issues that must be considered when developing PRO-PM. As noted above, many factors may affect the appropriateness of a PRO-PM:

- Is the performance measure reliable at the provider level?
    - Has it been tested to examine the random error associated with the provider-level unit of analysis separately from the individual-level error.
    - Is the sample size adequate for providing robust results?
    - Have any adjustments been made to reduce random error which may lead to misleading results?
- Is the performance measure construct valid? Does it allow you to make inferences about the organization?
    - Are the results statistically significant?

    - Are they clinically meaningful?
    - Are these PM important to the patient?
- How are the constructs being measured?
    - When is a change score more appropriate than a threshold value for an expected outcome?
    - How should the performance score be defined and what affects these decisions?
    - How has validity been tested? Does it meet face validity as an acceptable clinical expectation, criterion validity which measures up to a “gold standard” not likely to be repeated in a clinical environment, or construct validity with measurable differences between groups who vary in their implementation of quality improvement initiatives.
- Patient preferences – how are these taken into account ?
- Risk adjustment is another major area of consideration.
    - The exact covariates depend on the outcome being measured and the factors that may affect those outcomes, independent of the treatment provided.
    - Controlling for test effects such as the sources of data, the methods used, and the mode of administration may be important factors that can affect outcome scores independent of individual-level scores.
    - The methodologies for adjusting for these different risk factors vary and again, depend on the question being examined, and to some extent the preferences of the research team.
- Threats to validity – what factors may affect the validity of the performance measure after controlling for all the factors above?
    - Instrument validity
    - Missing data
    - Inadequate risk adjustment

The science of outcomes-based performance measurement is still relatively young. Patient-reported outcomes are even new forms of performance measurement. Being aware of the vagaries that may affect measurement reliability, even after ensuring a measure is valid and making decisions to select an appropriate approach for scoring performance and measuring outcomes is complex. Few answers are right or wrong; the best approach will likely depend on the goal of the performance monitoring. Meeting internal quality improvement goals should take into account patient preferences whereas outcome measures designed to meet regulatory requirements of ensuring at least a minimal level of quality may weigh these factors differently.

Processes that encourage the use of PROMs in daily clinical practice are needed so that best practices for data collection that occur within the clinical workflow can be identified. This may mean using process performance measures that are tied to PROM data collection as a starting point. More widespread use of PROMs in clinical practice will allow validity testing of PRO-based performance measures beyond face validity. Some of these issues can be addressed through the advent of eMeasures. As efforts move forward to develop more standardized EHRs, the standardized items are being incorporated into clinical practice. Efforts such as those spearheaded by the Office of the National Coordinator and the Centers for Medicare & Medicaid Services will help lay the groundwork for incorporating these items into daily treatment and workflow processes.

Moving the patient's voice into clinical practice is key for the future of person-centered health care. As noted by the Institute of Medicine in *Crossing the Quality Chasm*, six of the ten recommendations for improving the quality of the healthcare system address direct involvement of patients in their care. Engaging the patient in the process of care, particularly by noting their perceived outcomes, is key to developing better outcomes. Much more work is needed in this area to develop a robust set of measures that include the patient's voice in determining whether good outcomes of care have been achieved.

References

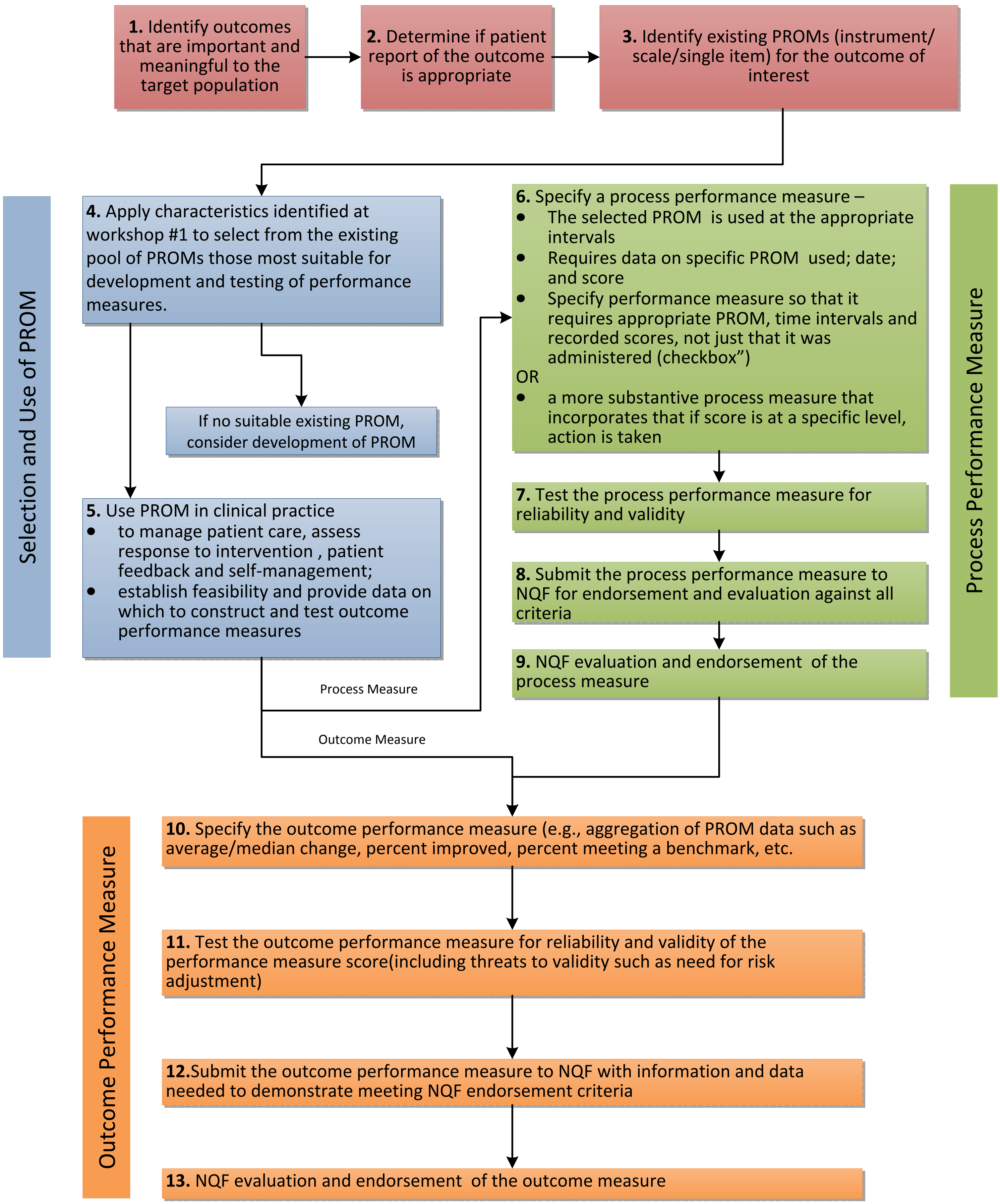
1. National Quality Forum. National voluntary consensus standards for patient outcomes. 2009
2. Cella D, Hahn EA, Jensen SE, Butt Z, Nowinski CJ, Rothrock NE. Methodological issues in the selection, administration ans use of patient-reported outcomes in performance measurment in health care settings. 2012
3. Kane RL, Radosevich, D. M. *Conducting health outcomes research*. Sudbury, MA: Jones & Bartlett Learning; 2011.
4. Iezonni L. *Risk adjustment for healthcare outcomes*. 2003.
5. McGlynn EA, Asch SM. Developing a clinical performance measure. *American journal of preventive medicine*. 1998;14:14-21
6. National Quality Forum. Guidance for measure testing and evaluating scientific acceptability of measure properties. 2011
7. Adams JL. The reliability of provider profiling: A tutorial. 2009
8. Adams JL, Mehrotra, Ateev, McGlynn, Elizabeth A. Estimating reliability and misclassification in physician profiling. 2010
9. Streiner DL, Norman GR. *Health measurement scales : A practical guide to their development and use*. Oxford ; New York: Oxford University Press; 2008.
10. Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: The importance of reliability adjustment. *Health services research*. 2010;45:1614-1629
11. Kao LS, Ghaferi AA, Ko CY, Dimick JB. Reliability of superficial surgical site infections as a hospital quality measure. *Journal of the American College of Surgeons*. 2011;213:231-235
12. Zaslavsky AM. Statistical issues in reporting quality data: Small samples and casemix variation. *International Journal for Quality in Health Care*. 2001;13:481-488
13. Kaplan SH, Griffith JL, Price LL, Pawlson LG, Greenfield S. Improving the reliability of physician performance assessment: Identifying the "physician effect" on quality and creating composite measures. *Medical care*. 2009;47:378-387
14. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA : the journal of the American Medical Association*. 1999;281:2098-2105
15. Roebroeck M, Harlaar J, Lankhorst G. The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Physical therapy*. 1993;73:386-395
16. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Physical therapy*. 1993;73:386-395; discussion 396-401
17. Austin PC. The reliability and validity of bayesian measures for hospital profiling: A monte carlo assessment. *J Statist Plann Inference*. 2005;128:109-122
18. Normand SL, Glickman ME, CA G. Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association*.92:803-814
19. Ash A, Fienberg SE, Louis TA, Normand S-LT, Stukel TA, Utts J. Statistical issues in assessing hospital performance. 2012. Available at: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf
20. Basch E. The missing voice of patients in drug-safety reporting. *The New England journal of medicine*. 2010;362:865-869
21. Tulsky DS, Kisala PA, Victorson D, Tate D, Heinemann AW, Amtmann D, Cella D. Developing a contemporary patient-reported outcomes measure for spinal cord injury. *Arch Phys Med Rehabil*. 2011;92:S44-51

22. Donabedian A. Evaluating the quality of medical care. 1966. *The Milbank quarterly*. 2005;83:691-729
23. US Department of Health and Human Services Food and Drug Administration. Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. 2009
24. Collins LM, Johnston MV. Analysis of stage-sequential change in rehabilitation research. *American Journal of Physical Medicine & Rehabilitation*. 1995;74:163-170
25. Stineman MG, Henry-Sanchez JT, Kurichi JE, Pan Q, Xie D, Saliba D, Zhang Z, Streim JE. Staging activity limitation and participation restriction in elderly community-dwelling persons according to difficulties in self-care and domestic life functioning. *American Journal of Physical Medicine & Rehabilitation*. 2012;91:126-140
26. Stineman MG, Maislin G, Fiedler RC, Granger CV. A prediction model for functional recovery in stroke. *Stroke*. 1997;28:550-556
27. Stineman MG, Ross RN, Fiedler R, Granger CV, Maislin G. Staging functional independence validity and applications. *Archives of Physical Medicine & Rehabilitation*. 2003;84:38-45
28. Stineman MG, Ross RN, Fiedler R, Granger CV, Maislin G. Functional independence staging: Conceptual foundation, face validity, and empirical derivation. *Archives of Physical Medicine & Rehabilitation*. 2003;84:29-37
29. Stineman MG, Xie D, Pan Q, Kurichi JE, Saliba D, Streim J. Activity of daily living staging, chronic health conditions, and perceived lack of home accessibility features for elderly people living in the community. *Journal of the American Geriatrics Society*. 2011;59:454-462
30. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank quarterly*. 2004;82:661-687
31. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the patient health questionnaire (phq-9): A meta-analysis. *CMAJ Canadian Medical Association Journal*. 2012;184:E191-196
32. Fitch K. *The rand/ucla appropriateness method user's manual*. Santa Monica: Rand; 2001.
33. Spertus JA, Eagle KA, Krumholz HM, Mitchell KR, Normand S-LT, American College of Cardiology/American Heart Association Task Force on Performance M. American college of cardiology and american heart association methodology for the selection and creation of performance measures for quantifying the quality of cardiovascular care. *Journal of the American College of Cardiology*. 2005;45:1147-1156
34. Hibbard J, Pawlson LG. Why not give consumers a framework for understanding quality? *Jt Comm J Qual Saf*. 2004;30:347-351
35. National Quality Forum. Csac guidance on quality performance measure construction. 2011
36. Skolarus LE, Sánchez BN, Morgenstern LB, Garcia NM, Smith MA, Brown DL, Lisabeth LD. Validity of proxies and correction for proxy use when evaluating social determinants of health in stroke patients. *Stroke; A Journal Of Cerebral Circulation*. 2010;41:510-515
37. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, Lenderking WR, Cella D, Basch E. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (pro) measures: Ispor epro good research practices task force report. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*. 2009;12:419-429
38. Saliba D, Buchanan, Joan. Development and validation of a revised nursing home assessment tool: Mds 3.0. 2008
39. Saliba D, DiFilippo S, Edenlen MO, Kroenke K, Buchanan J, Streim J. Testing the phq-9 interview and observational versions (phq-9 ov) for mds 3.0. *JAMDA*. 2012;in press

40. Agnihotri K, Awasthi S, Singh U, Chandra H, Thakur S. A study of concordance between adolescent self-report and parent-proxy report of health-related quality of life in school-going adolescents. *Journal of psychosomatic research*. 2010;69:525-532
41. Chang PC, Yeh CH. Agreement between child self-report and parent proxy-report to evaluate quality of life in children with cancer. *Psycho-oncology*. 2005;14:125-134
42. Matza LS, Secnik K, Rentz AM, Mannix S, Sallee FR, Gilbert D, Revicki DA. Assessment of health state utilities for attention-deficit/hyperactivity disorder in children using parent proxy report. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*. 2005;14:735-747
43. Varni JW, Limbers CA, Burwinkle TM. Parent proxy-report of their children's health-related quality of life: An analysis of 13,878 parents' reliability and validity across age subgroups using the pedsql 4.0 generic core scales. *Health and quality of life outcomes*. 2007;5:2
44. Varni JW, Stucky BD, Thissen D, Dewitt EM, Irwin DE, Lai J-S, Yeatts K, Dewalt DA. Promis pediatric pain interference scale: An item response theory analysis of the pediatric pain item bank. *The Journal Of Pain: Official Journal Of The American Pain Society*. 2010;11:1109-1119
45. Sagheri D, Wiater A, Steffen P, Owens JA. Applying principles of good practice for translation and cross-cultural adaptation of sleep-screening instruments in children. *Behavioral sleep medicine*. 2010;8:151-156
46. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, Erikson P. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (pro) measures: Report of the ispor task force for translation and cultural adaptation. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*. 2005;8:94-104
47. Agoritsas T, Lubbeke A, Schiesari L, Perneger TV. Assessment of patients' tendency to give a positive or negative rating to healthcare. *Quality & safety in health care*. 2009;18:374-379
48. Johnson TP, Wislar JS. Response rates and nonresponse errors in surveys. *JAMA : the journal of the American Medical Association*. 2012;307:1805-1806
49. Casarett D, Smith D, Breslin S, Richardson D. Does nonresponse bias the results of retrospective surveys of end-of-life care? *Journal of the American Geriatrics Society*. 2010;58:2381-2386
50. Klein DJ, Elliott MN, Haviland AM, Saliba D, Burkhart Q, Edwards C, Zaslavsky AM. Understanding nonresponse to the 2007 medicare cahps survey. *Gerontologist*. 2011;51:843-855
51. Dillman DA, Smythe HD, Christian LM. *Internet, mail and mixed_mode surveys: The tailored design method*. Hoboken, NJ: John Wiley and Sons, Inc; 2009.
52. Elliott MN, Beckett MK, Chong K, Hambarsoomians K, Hays RD. How do proxy responses and proxy-assisted responses differ from what medicare beneficiaries might have reported about their health care? *Health services research*. 2008;43:833-848

Pathway from Individual-level PRO instrument/measure to NQF-endorsed Performance Measure

September 7, 2012



1. Identify outcomes that are important and meaningful to the target population

2. Determine if patient report of the outcome is appropriate

3. Identify existing PROMs (instrument/ scale/single item) for the outcome of interest

Selection and Use of PROM

4. Apply characteristics identified at workshop #1 to select from the existing pool of PROMs those most suitable for development and testing of performance measures.

If no suitable existing PROM, consider development of PROM

5. Use PROM in clinical practice

- to manage patient care, assess response to intervention , patient feedback and self-management;
- establish feasibility and provide data on which to construct and test outcome performance measures

Process Measure

Outcome Measure

Process Performance Measure

6. Specify a process performance measure –

- The selected PROM is used at the appropriate intervals
- Requires data on specific PROM used; date; and score
- Specify performance measure so that it requires appropriate PROM, time intervals and recorded scores, not just that it was administered (checkbox")

OR

- a more substantive process measure that incorporates that if score is at a specific level, action is taken

7. Test the process performance measure for reliability and validity

8. Submit the process performance measure to NQF for endorsement and evaluation against all criteria

9. NQF evaluation and endorsement of the process measure

Outcome Performance Measure

10. Specify the outcome performance measure (e.g., aggregation of PROM data such as average/median change, percent improved, percent meeting a benchmark, etc.

11. Test the outcome performance measure for reliability and validity of the performance measure score(including threats to validity such as need for risk adjustment)

12. Submit the outcome performance measure to NQF with information and data needed to demonstrate meeting NQF endorsement criteria

13. NQF evaluation and endorsement of the outcome measure

PATHWAY NOTES – Correspond to Pathway Elements on Page 1

NQF Criteria
Importance to Measure and Report

If patient/person is not the best source of information for the outcome, then explore clinical data and measurement

PRO refers to the concept
PROM refers to the instrument, scale, or single-item to measure the PRO concept
PRO-PM refers to PRO-based performance measure
Many PROMs developed and tested (reliability, validity, responsiveness, identification of meaningful differences, etc.) primarily for research

NQF Criteria
Importance to Measure and Report, 1c. Evidence - responsive to clinical intervention

Scientific Acceptability of Measure Properties:
2a. Reliability and 2b. Validity
Reliability and validity of data elements used in performance measure (i.e., PROM data)

Characteristics for Selecting PROMs Identified in Commissioned Paper (Table 3)
1. Conceptual and Measurement Model Documented
2. Reliability
2a. *Internal consistency (multi-item scales)*
2b. *Reproducibility (stability over time)*
3. Validity
3a. Content Validity
3b. Construct and Criterion-related Validity
3c. Responsiveness
4. Interpretability of Scores 8
5. Burden
6. Alternatives modes and methods of administration
7. Cultural and language adaptations
8. Electronic health record (EHR) capability

Additional Characteristics from workshop:
Meaningful · Actionable · Able to facilitate shared decision-making · Implementable

NQF Criteria
Usability and Use
Feasibility

Process measure considered an interim step to encourage use and obtain data and experience so that outcome performance measure could be developed, tested, and endorsed

NQF Criteria
2a.1 Precise specification
2b.1 Specifications consistent with evidence

Should be specified so that data can be used to construct and test future outcome performance measures

Should be specified so that it is more than a "checkbox" – "checkbox" measures generally do not pass Importance to Measure and Report because not proximal to desired outcomes; doing an assessment is first step but far from sufficient to influence outcomes

NQF Criteria
2a2 Reliability testing
2b2 Validity testing

Is using a reliable and valid PROM sufficient demonstration of reliability and validity at the data element level? Is testing at the level of the performance measure needed?

NQF Criteria
Importance to Measure and Report
Scientific Acceptability of Measure Properties
Usability and Use
Feasibility

Does endorsement of performance measure increase use and provide more data/ experience to develop and test outcome performance measure?
Or does it divert focus and resources from outcome?

NQF Criteria
2a.1 Precise specification
2b.1 Specifications consistent with evidence

NQF Criteria
2a2. Reliability testing of performance measure score, e.g., signal-to-noise analysis (or is data element sufficient?)
2b2. Validity testing of the performance measure score i.e., can make correct conclusions about quality of care (or is data element sufficient?)
2b3. Exclusions justified
2b4. Differences in case-mix (is risk adjustment needed, adequate?)
2b5. Performance measure score discriminates among the accountable entities
2b6. Comparability of different data sources/methods

NQF Criteria
Importance to Measure and Report
Scientific Acceptability of Measure Properties
Feasibility
Usability and Use

NATIONAL QUALITY FORUM

During the first workshop the Expert Panel discussed high leverage characteristics for identifying PROMs most suitable for development and testing of performance measures. Workshop participants agreed the psychometric properties captured in the attached table derived from the commissioned paper written by David Cella and team were considered as baseline, but also offered additional guideposts for consideration. NQF staff compiled this feedback and the Expert Panel was offered an opportunity to provide further input through a survey. Further exploration and refinement of these characteristics will take place at the second workshop and their relationship to the NQF endorsement evaluation criteria for which they are mutually reinforcing.

Below is a distillation of the proposed edits received from the survey. Redlines are included to depict specific edits.

- Meaningful to ~~persons~~ patients, ~~and~~ their families and caregivers, ~~– as well as clinicians and~~ and to ~~other~~ health professionals who serve them. Meaningfulness encompasses the relevance and degree of importance of the concepts measured by the PROM from the perspective of each of these stakeholders. Among meaningful concepts for PROMs to ~~adequately~~ capture are: ~~the~~ impact of health related quality of life ~~(including functional status)~~, symptom and symptom burden, experience with care, ~~or of~~ a health-related behavior on the patient, or community-based health services and supports.

- Actionable with ~~evidence-based justification~~ criteria -for selection based on evidence or strong professional consensus that data gathered on the outcome ~~that~~ leads to improvement in heath, care quality, or services/supports received by key end users (e.g., persons, providers, systems). ~~including persons, providers and systems~~ ~~.~~

- Able to facilitate shared decision-making —the measure (PROM) will ~~including~~ engag~~ing~~e patients in their ~~own~~ preferred self-management and goal attainment ~~aligned with their preferences~~ (e.g., by identifying outcomes important to them and ~~;~~ involving them in measure development and testing~~;~~ in ways ~~assessing~~that are culturally/ linguistically adaptab~~ility~~le~~)~~ while being ~~flexible enough~~ sufficiently standardized to permit ~~to~~ aggregat~~e~~ion or roll up ~~for~~to a population or accountable entity.

- Implementable taking into account burden to the person, provider, and system including but not limited to: cost barriers to the use of proprietary tools or measures; ease of fielding; potential for unintended consequences (e.g., gaming or adverse selection); shown to be successfully integrated into routine clinical practice and into patient's daily lives; ~~measures that are disparities sensitive;~~ and adaptability to electronic or other alternate formats.

Table 4[1] (Previously Table 3). Important characteristics and best practices to evaluate and select PROs for use in performance measures[274,279]

Authors of 1st paper: David Cella, Ph.D., Elizabeth A. Hahn, M.A., Sally E. Jensen, Ph.D., Zeeshan Butt, Ph.D., Cindy J. Nowinski, M.D., Ph.D., Nan Rothrock, Ph.D

	Characteristic	**Specific issues to address for performance measures**	**Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[349] for use in hip arthroplasty**
1.	**Conceptual and Measurement Model**		
	A PRO measure should have documentation defining and describing the concept(s) included and the intended population(s) for use.	• Target PRO concept should be a high priority for the health care system and patients. Patient engagement should define what is an important concept to the patients. • Target PRO concept must be actionable in response to the healthcare intervention.	• Factorial validity of the physical function and pain subscales has been inadequate.[350]
	There should be documentation of how the concept(s) are organized into a measurement model, including evidence for the dimensionality of the measure, how items relate to each measured concept, and the relationship among concepts.		
2.	**Reliability**		
	The degree to which an instrument is free from random error.		
2a.	***Internal consistency*** *(multi-item scales)*	Classical Test Theory (CTT): ▪ reliability estimate ≥ 0.70 for group-level purposes ▪ reliability estimate ≥ 0.90 for individual-level purposes Item Response Theory: • item information curves that demonstrate precision [176] • a formula can be applied to estimate CTT reliability	• Cronbach alphas for the three subscales range from 0.86 to 0.98.[351-353]
2b.	***Reproducibility*** *(stability over time)* ▪ type of test-retest estimate depends on the response scale (dichotomous, nominal ordinal, interval, ratio)		• Test-retest reliability has been adequate for the pain and physical function subscales, but less adequate for the stiffness subscale.[353]
3.	**Validity**		

[1] This table is adapted from recommendations contained within a report from the Scientific Advisory Committee of the Medical Outcomes Trust and a report submitted to the PCORI Methodology Committee. The recommendations from these sources have been adapted to enhance relevance to PRO selection for performance measurement.

	Characteristic	**Specific issues to address for performance measures**	**Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[349] for use in hip arthroplasty**
	The degree to which the instrument reflects what it is supposed to measure.	• There are a limited number of PRO instruments that have been validated for performance measurement. • PRO instruments should include questions that are patient-centered.	
3a.	***Content Validity***		
	The extent to which a measure samples a representative range of the content.		
	A PRO measure should have evidence supporting its content validity, including evidence that patients and/or experts consider the content of the PRO measure relevant and comprehensive for the concept, population, and aim of the measurement application.		• Development involved expert clinician input, and survey input from patients,[354] as well as a review of existing measures.
	Documentation of qualitative and/or quantitative methods used to solicit and confirm attributes (i.e., concepts measured by the items) of the PRO relevant to the measurement application.		
	Documentation of the characteristics of participants included in the evaluation (e.g., race/ethnicity, culture, age, socio-economic status, literacy).		
	Documentation of sources from which items were derived, modified, and prioritized during the PRO measure development process.		
	Justification for the recall period for the measurement application.		
3b.	***Construct and Criterion-related Validity***		
	A PRO measure should have evidence supporting its construct validity, including: • documentation of empirical findings that support predefined hypotheses on the expected associations among measures similar or dissimilar to the measured PRO • documentation of empirical findings that support predefined hypotheses of the expected differences in scores between "known" groups		• Patient ratings of satisfaction with arthroplasty were correlated with WOMAC scores in the expected direction.[22,355,356]
	A PRO measure should have evidence that shows the extent to which scores of the instrument are related to a criterion measure.		
3c.	***Responsiveness***		
	A PRO measure for use in longitudinal initiatives should have evidence of responsiveness, including empirical evidence of changes in scores consistent with predefined hypotheses regarding changes in the target population.	• If a PRO measure has cross-sectional data that provides sufficient evidence in regard to the reliability (internal consistency), content validity, and construct validity but has no data yet on responsiveness over time (i.e., ability of a PRO measure to detect changes in the construct being measured over time),	• Demonstrates adequate responsiveness and ability to detect change in response to clinical intervention.[357]

	Characteristic	Specific issues to address for performance measures	Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[349] for use in hip arthroplasty
		would you accept use of the PRO measure to provide valid data over time in a longitudinal study if no other PRO measure was available?	
		• Important to emphasize responsiveness because there is an expectation of consequences. Need to be able to demonstrate responsiveness if action is to be taken.	
		• PRO must be sensitive to detect change in response to the specific healthcare intervention	
4.	**Interpretability of Scores**		
	A PRO measure should have documentation to support interpretation of scores, including: • what low and high scores represent for the measured concept • representative mean(s) and standard deviation(s) in the reference population • guidance on the minimally important difference in scores between groups and/or over time that can be considered meaningful from the patient and/or clinical perspective	• If different PROs are used, it is important to establish a link or cross-walk between them. • Because the criteria for assessing clinically important change in individuals does not directly translate to evaluating clinically important group differences, [322] a useful strategy is to calculate the proportion of patients who experience a clinically significant change[266,322]	• Availability of population-based, age- and gender-normative values[358] • Availability of minimal clinically important improvement values[359] • Can be translated into a utility score for use in economic and accountability evaluations[360]
5.	**Burden**		
	The time, effort, and other demands on the respondent and the administrator.	• In a busy clinic setting, PRO assessment should be as brief as possible, and reporting should be done in real-time. • Patient engagement should inform what constitutes "burden."	• Short form available[361] • Average time to complete mobile phone WOMAC = 4.8 minutes[362]
6.	**Alternatives modes and methods of administration**	• The use of multiple modes and methods can be useful for diverse populations. However, there should be evidence regarding their equivalence.	• Validated mobile phone and touchscreen based platforms[363,364]
7.	**Cultural and language adaptations**	• The mode, method and question wording must yield equivalent estimates of PRO measures.	• Available in over 65 languages[365]
		Critical features:	▪ Electronic data

	Characteristic	Specific issues to address for performance measures	Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[349] for use in hip arthroplasty
8.	**Electronic health records (EHR)**	▪ interoperability ▪ automated, real-time measurement and reporting ▪ sophisticated analytic capacities	capture may allow for integration within EHR[362]

NATIONAL QUALITY FORUM

Measure Evaluation Criteria
January 2011

Conditions for Consideration

Several conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards. **If any of the conditions are not met, the measure will not be accepted for consideration.**

A. The measure is in the public domain or a measure steward agreement is signed.

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

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

D. The measure is fully specified and tested for reliability and validity.[1]

E. The measure developer/steward attests that harmonization with related measures and issues with competing measures have been considered and addressed, as appropriate.

F. The requested measure submission information is complete and responsive to the questions so that all the information needed to evaluate all criteria is provided.

Note

1. A measure that has not been tested for reliability and validity is only potentially eligible for time-limited endorsement if all of the following conditions are met: 1) the measure topic is not addressed by an endorsed measure; 2) it is relevant to a critical timeline (e.g., legislative mandate) for implementing endorsed measures; 3) the measure is not complex (requiring risk adjustment or a composite); and 4) the measure steward verifies that testing will be completed within 12 months of endorsement.

Criteria for Evaluation

If all conditions for consideration are met, candidate measures are evaluated for their suitability based on four sets of standardized criteria in the following order: *Importance to Measure and Report*, *Scientific Acceptability of Measure Properties*, *Usability*, and *Feasibility*. Not all acceptable measures will be equally strong among each set of criteria. The assessment of each criterion is a matter of degree. However, if a measure is not judged to have met minimum requirements for *Importance to Measure and Report* or *Scientific Acceptability of Measure Properties*, it cannot be recommended for endorsement and will not be evaluated against the remaining criteria.

1. Impact, Opportunity, Evidence—Importance to Measure and Report: Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-impact aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to*** *meet all three subcriteria to pass this criterion and be evaluated against the remaining criteria.*

1a. High Impact

The measure focus addresses:

- a specific national health goal/priority identified by 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. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data[2] demonstrating considerable variation, or overall less-than-optimal performance, in the quality of care across providers and/or population groups (disparities in care).

AND

1c. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- 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 OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

2. 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.

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

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 multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

2. Reliability and Validity—Scientific Acceptability of Measure Properties: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. *Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.*

2a. Reliability

2a1. The measure is well defined and precisely specified[7] so it can be implemented consistently within and across organizations and allow for comparability. EHR measure specifications are based on the quality data model (QDM).[8]

2a2. Reliability testing[9] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

2b. Validity

2b1. The measure specifications[7] are consistent with the evidence presented to support the focus of measurement under criterion 1c. The measure is specified to capture the most inclusive target population indicated by the evidence, and exclusions are supported by the evidence.

2b2. Validity testing[10] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion;[11]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).[12]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care;[13,14] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful[15] differences in performance;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2c. Disparities

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.

Notes

7. Measure specifications include the target population (denominator) to whom the measure applies, identification of

those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window, exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, scoring/computation.
8. EHR measure specifications include data type from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting.
9. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements 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 of the measure score addresses precision of measurement (e.g., signal-to-noise).
10. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
11. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
12. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
13. Risk factors that influence outcomes should not be specified as exclusions.
14. 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, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.
15. 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 percent v. 75 percent) 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 less-than-optimal performance may not demonstrate much variability across providers.

Note: This criterion revised February 2012 and will be replaced Fall 2012 - see end of document
3. Usability: Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) can understand the results of the measure and find them useful for decisionmaking.

3a. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for public reporting (e.g., focus group, cognitive testing) or rationale;

AND

3b. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for informing quality improvement[16] (e.g., quality improvement initiatives) or rationale.

Note
16. 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.

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

4a. For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

4b. The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

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

4d. 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).

Note
[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.

5. Comparison to Related or Competing Measures
If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5a. The measure specifications are harmonized[18] with related measures;

OR

the differences in specifications are justified.

5b. The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

multiple measures are justified.

Note
[18] Measure harmonization refers to the standardization of specifications for related measures with the same measure focus (e.g., *influenza immunization* of patients in hospitals or nursing homes); related measures with 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 justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, 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.

Evaluation Criteria for Usability and Use (Feb 2012) – Will replace prior criterion Fall 2012

Condition for Consideration

C. The intended use of the measure includes both accountability applications[1] and performance improvement to achieve high-quality, efficient healthcare.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement[2] to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency[3]

Performance results are used in at least one accountability application[1] within three years after initial endorsement and are publicly reported[3] within six years after initial endorsement (or the data on performance results are available).[4] If not in use at the time of initial endorsement, then a credible plan[5] for implementation within the specified timeframes is provided.

AND

4b. Improvement[6]

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.[6] If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

AND

4c. The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Criteria Notes

1. Accountability applications are the use of performance results about identifiable, accountable entities to make judgments and decisions as a consequence of performance, such as reward, recognition, punishment, payment, or selection (e.g., public reporting, accreditation, licensure, professional certification, health information technology incentives, performance-based payment, network inclusion/exclusion). Selection is the use of performance results to make or affirm choices regarding providers of healthcare or health plans.
2. 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.
3. Transparency is the extent to which performance results about identifiable, accountable entities are *disclosed and available* outside of the organizations or practices whose performance is measured. Maximal transparency is achieved with public reporting defined as making comparative performance results about identifiable, accountable entities freely available (or at nominal cost) to the public at large (generally on a public website). *At a minimum, the data on performance results about identifiable, accountable entities are available to the public (e.g., unformatted database)*. The capability to verify the performance results adds substantially to transparency.
4. This guidance is not intended to be construed as favoring measures developed by organizations that are able to implement their own measures (such as government agencies or accrediting organizations) over equally strong measures developed by organizations that may not be able to do so (such as researchers, consultants, or academics). Accordingly, measure developers may request a longer timeframe with appropriate explanation and justification.
5. Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.
6. Demonstrated progress toward achieving the goal of high-quality, efficient healthcare includes evidence of improved performance and/or increased numbers of individuals receiving high-quality healthcare. Exceptions may be considered with appropriate explanation and justification.

NATIONAL QUALITY FORUM

Patient-Reported Outcomes Expert Panel

EXPERT PANEL MEMBERS

Dr. Richard Bankowitz, MBA, MD, FACP
Chief Medical Officer, Premier healthcare alliance

Richard Bankowitz, MD, MBA, FACP is currently serving as Chief Medical Officer of Premier healthcare alliance. Dr. Bankowitz engages physicians, provides leadership and ensures that Premier continues to deliver value to its clinician constituency. Dr. Bankowitz previously served as VP and medical director for Premier Healthcare Informatics. A board-certified internist and a medical informaticist, Dr. Bankowitz has devoted his career to improving healthcare quality at the national level by promoting rigorous, data-driven approaches to quality improvement. He began his career at the University of Pittsburgh School of Medicine as an assistant professor of medicine and medical informatics and served as the architect of the University HealthSystem Consortium.

Dr. Ethan Basch, MD, MSc
Associate Attending Physician and Outcomes Research Scientist- Memorial Sloan-Kettering Cancer Center

Dr. Ethan Basch is an oncologist and outcomes researcher at Memorial Sloan-Kettering Cancer Center who directs a program on patient-reported outcomes, clinical informatics, comparative effectiveness and product safety evaluation. He leads the National Cancer Institute's PRO-CTCAE initiative to develop a standardized patient-centered approach to adverse event reporting in clinical trials. He is a member of the PCORI Methodology Committee and chairs the Patient-Centeredness Workgroup. He is immediate past Chair of the American Society of Clinical Oncology Clinical Practice Guidelines Committee, member of the Comparative Effectiveness Research Task Force and liaison to the Quality of Care Committee. Dr. Basch received his MD from Harvard.

Dr. Jim Bellows, PhD, MPH
Senior Director, Evaluation and Analytics- Kaiser Permanente Care Management Institute (CMI)

Jim Bellows is Senior Director, Evaluation and Analytics in Kaiser Permanente's Care Management Institute. Dr. Bellows leads an Evaluation and Analytics staff with expertise in metrics development, analytics, and quantitative and qualitative evaluation, and with accountability for developing and producing performance metrics, identifying specific population care practices that contribute to superior performance, and evaluating the impact of quality improvement initiatives as they mature. Dr. Bellows shares responsibility for building collaborations that apply the capabilities of KP's externally-funded research investigators to clinical and operational challenges of strategic importance within KP.

Dr. Patricia Flatley Brennan, RN, PhD
Professor, School of Nursing and College of Engineering, University of Wisconsin

Dr. Brennan is the Lillian L. Moehlman Bascom Professor at the School of Nursing and College of Engineering, University of Wisconsin-Madison. Dr. Brennan received a Masters of Science in Nursing from the University of Pennsylvania and a PhD in Industrial Engineering from the University of Wisconsin-Madison. Following seven years of clinical practice in critical care nursing and psychiatric nursing, Dr. Brennan held several academic positions. She developed the ComputerLink, an electronic network designed to reduce isolation and improve self-care among home care patients and directed HeartCare, a WWW-based tailored information and communication service that helped home-dwelling cardiac patients recover faster, and with fewer symptoms.

Ms. Laurie Burke, RN
Associate Director for Study Endpoints and Labeling in the Center for Drug Evaluation and Research, Food and Drug Administration

Ms. Burke is an advocate for the development of good measurement practices and leads in the development of regulatory policy for outcome assessments to support claims in labeling. She has led many FDA-wide initiatives including the publication of FDA's Patient Reported Outcome guidance and has authored numerous white papers through her involvement in FDA working groups and professional associations. The Study Endpoints Team identifies best measurement practices and works with all FDA medical product reviewers to determine whether clinical outcome assessments are well-defined and reliable with respect to their context of use in the support of medical product development, labeling, and promotion.

Ms. Joyce Dubow, MUP
Senior Director, Health Care Reform- AARP

Ms. Dubow is a Senior Advisor in AARP's Office of Policy and Strategy where she has responsibility for a broad portfolio, including health care quality, measurement, public reporting, patient decision making, HIT, and related issues. Before joining this office, she was Associate Director in AARP's Public Policy Institute where I had responsibility for public policy research and analysis. In a "former life," she was executive vice-president of the Georgetown University Community Health Plan, a university-sponsored prepaid group practice plan. Ms. Dubow also served as the Director of Policy and Legislation in the federal Office of Health Maintenance Organizations.

Ms. Jennifer Eames-Huff, MPH
Director, Consumer-Purchaser Disclosure Project- Pacific Business Group on Health

Ms. Eames Huff is Director for the Consumer-Purchaser Disclosure Project, which is a group of leading employer, consumer, and labor organizations improving health care quality and affordability by advancing public reporting of provider performance information so it can be used for improvement, consumer choice, and payment. Ms. Huff brings over fifteen years experience working in the arena of health care performance measurement to the project. Prior to joining PBGH Ms. Huff was a Health Economist at Genentech. Before that, she was a Program Officer at the California HealthCare Foundation. Ms. Huff earned a BA with Honors from Wellesley College and an MPH in Health Policy and Management from University of California at Berkeley.

Dr. Stephan Fihn, MD, MPH
Director, Office of Analytics and Business Intelligence, Veterans Health Administration

Dr. Stephan Fihn is a general internist and health services researcher at VA Puget Sound Health Care System and the University of Washington in Seattle. Until recently, he served as Director of the Northwest VA Health Services Research & Development Center of Excellence at VAPSHCS. He now serves as Director, Office of Analytics and Business Intelligence for Veterans Health Administration. His office will provide comprehensive analytic and business intelligence support to all of VHA. He also serves as Head of the Division of General Internal Medicine at the University of Washington. His research interests relate to developing strategies for improving the efficiency and quality of primary medical care and understanding the epidemiology of common ambulatory problems.

Dr. Floyd Jackson Fowler, Jr., PhD
Senior Scientific Advisor and Past President- Foundation for Informed Medical Decision Making

Floyd J Fowler Jr. is a Senior Scientific Advisor to the Foundation for Informed Medical Decision Making. He served as President of the Foundation from 2002-2009. He has also been a Senior Research Fellow at the Center for Survey Research, UMass Boston since 1971, and he served as Director of the Center for 14 years. Dr. Fowler is a social scientist whose special expertise is survey methodology. He is the author (or co-author) of four widely used books on survey research methods. He also has been a major contributor to research on patient outcomes and on how patients are affected by the treatments they receive. Dr. Fowler received a BA degree in English from Wesleyan University and a Ph.D. is Social Psychology from the University of Michigan.

Dr. Lori Frank, PhD
Director, Engagement Research- Patient Centered Outcomes Research Institute

Lori Frank, PhD, has worked as a PRO researcher for over 15 years. At PCORI Dr. Frank's work focuses on the patient perspective on comparative effectiveness research. As Executive Director of the Center for Health Outcomes Research at United BioSource she led multiple PRO development and psychometric evaluation studies, and initiated and led the Cognition Initiative, now part of the Critical Path Institute PRO Consortium. Her published work includes both qualitative and quantitative studies of PROs. She serves on the Memory Screening Advisory Board of the Alzheimer's Foundation of America and has also served on the Center for Trauma and the Community of Georgetown University Department of Psychiatry.

Dr. Theodore Ganiats, MD
Professor- University of California San Diego

Theodore G. Ganiats, MD, is Professor of Family and Preventive Medicine at the University of California San Diego (UCSD) School of Medicine and the Executive Director of the UCSD Health Services Research Center. Dr. Ganiats' research interests involve outcomes research, focusing on both applied and theoretical aspects of quality of life assessment (an important patient-reported outcome) and cost-effectiveness analysis. He has co-chaired or been a member of over fifty national systemic reviews, clinical practice guidelines (often as a methodology consultant) and performance measurement panels, including NQF heart failure and diabetes panels. He remains clinically active, giving him the additional perspective of the practicing clinician.

Dr. Kate Goodrich, MD
Senior Technical Advisor to the Director of the Office of Clinical Standards and Quality and Chief Medical Officer, Centers for Medicare and Medicaid Services

Dr. Goodrich earned her M.D. from Louisiana State University Medical Center. She completed her residency in Internal Medicine at George Washington University Medical Center, followed by a year as Chief Medical Resident. She joined the faculty of GWUMC as a hospitalist in the Department of Medicine. A new Division of Hospital Medicine was created in 2005, and Dr. Goodrich was appointed Division Director. From 2003-2008 she served as Chair of the Institutional Review Board at GWUMC. She also took a position as Medical Officer at the Department of Health and Human Services in the office of the Assistant Secretary for Planning and Evaluation. Dr. Goodrich continues to practice clinical medicine as a hospitalist at George Washington University Hospital.

Dr. Judith Hibbard, DrPH
Professor Emerita and Senior Researcher, Institute for Policy Research and Innovation, University of Oregon

Dr. Hibbard has focused her research on 1) how the presentation of quality data affects consumers' use of quality information in decision-making, 2) how health literacy affects choices, 3) measuring patient engagement and activation, and 4) whether public reporting stimulates quality improvement. She has led over a dozen studies using both qualitative and quantitative approaches. She has examined how to present quality data to highly vulnerable populations, such as Medicare beneficiaries, patients with chronic illnesses, and patients with low numeracy levels. One of the most important facets of her work is elaborating whether patients become more engaged in their own health in response to information.

Dr. Dennis Kaldenberg, PhD
Chief Scientist, Senior Vice President- Press Ganey Associates

As Chief Scientist, Dennis Kaldenberg provides leadership to the areas of research and analytics including such issues as data integration, information collection protocols and the accurate and useful dissemination of information. During his tenure he has been instrumental in the creation and revision of many tools to measure patient experience, patient satisfaction, and other patient reported outcomes. Dr. Kaldenberg has written on a variety of topics related to patient satisfaction, health care service delivery, health care professionals, and research methods. He has presented at the national meetings of numerous health care professional associations, Dr. Kaldenberg received his Ph.D. from Iowa State University with a specialization in research methods.

Dr. Irene L. Katzan, MD, MS
Director, Neurological Institute Center for Outcomes Research & Evaluation- Cleveland Clinic

Irene Katzan MD, MS is a board-certified vascular neurologist and health services researcher at Cleveland Clinic. She is Director of the Neurological Institute Center for Outcomes Research and Evaluation and the Knowledge Program, a technology initiative to harness electronic clinical information for research and patient care. Dr. Katzan is also a senior researcher at Case's Center for Health Care Research & Policy. She has a background in evaluating outcomes of care and modifying systems to optimize patient management in multiple settings. She is actively involved in regional stroke care initiatives and is the lead physician of the Ohio Coverdell Stroke Registry, a quality initiative of Centers for Disease Control & Prevention.

Dr. Lewis Kazis, Sc.D
Professor, Health Policy and Management- Boston University School of Public Health

Dr. Kazis is Professor of Health Policy and Management at Boston University School of Public Health. He has published well over 150 peer reviewed publications including those involving PROs. Dr. Kazis was the recipient of the Research Career Scientist Award from the VA for almost a decade. Dr. Kazis served as a special consultant to the Office of Quality and Performance in the VA and a consultant to the Centers for Medicare and Medicaid Services for the evaluation of the Medicare Advantage Program. Dr. Kazis is the principal developer of the Veterans RAND 36 and 12 Item Health Survey's (VR-36/12). The VR-12 has been adopted by the Veterans Administration historically and by the CMS Medicare Advantage Program.

Dr. Uma Kotagal, M.B.B.S, MSc
Senior Vice President for Safety, Quality and Transformation and Executive Director of the James M. Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center

Dr. Uma Kotagal is the Senior Vice President for Safety, Quality and Transformation and Executive Director of the James M. Anderson Center for Health Systems Excellence at Cincinnati Children's Hospital Medical Center. Dr. Kotagal is Chair of the Quality Steering Team of the Ohio Children's Hospital Association, member of the Advisory Committee of the Toronto Patient Safety Center, Associate Editor for the BMJ Quality and Safety, and is a member of the Institute of Medicine. Dr. Kotagal holds a MS in Epidemiology from Harvard

University-School of Public Health and a Bachelors of Medicine, Surgery from Grant Medical College, Mumbai, India.

Dr. Kevin Larsen, MD
Medical Director of Meaningful Use, Office of the National Coordinator

Kevin L. Larsen, MD is Medical Director of Meaningful Use at the Office of the National Coordinator for Health IT. In that role he is responsible for coordinating the clinical quality measures for Meaningful Use Certification and overseas the development of the Population Health Tool http://projectpophealth.org. Prior to working for the federal government he was Chief Medical Informatics Officer and Associate Medical Director at Hennepin County Medical Center in Minneapolis, Minnesota. He is also an Associate Professor of Medicine at the University of Minnesota. Dr. Larsen graduated from the University of Minnesota Medical School and was a resident and chief medical resident at Hennepin County Medical Center. He is a general internist and teacher in the medical school and residency programs. His research includes health care financing for people living in poverty, computer systems to support clinical decision making, and health literacy. In Minneapolis he was also the Medical Director for the Center for Urban Health, a hospital, community collaboration to eliminate health disparities. He served on a number of state and national committees in informatics, data standards and health IT.

Dr. Kathleen Lohr, PhD
Distinguished Fellow, RTI International

Kathleen N. Lohr, PhD, Distinguished Fellow at RTI International, was the founding director of the RTI–UNC Evidence-based Practice Center; recent projects involve systematic or comparative effectiveness reviews, “content” and “readability” guidance for AHRQ reviews, and EPC methods projects. Dr. Lohr was the founding Editor-in-Chief of RTI Press (www.rti.org/RTIPress); she was Associate Editor of Quality of Life Research and is on the editorial board of Comparative Effectiveness Research. She received (2005) the Avedis Donabedian Outcomes Research Lifetime Achievement Award from the International Society of Pharmacoeconomics & Outcomes Research. She was a member of AHRQ’s National Advisory Council (2008-2010).

Dr. Elizabeth Mort, MD
Associate Chief Medical Officer, Senior Vice President Quality and Safety, Massachusetts General Hospital

Dr. Mort oversees data collection and external reporting for public reporting for MGH. She sits on the hospital’s Quality Executive Committee and the Board Subcommittee on Quality which set the quality and safety agenda for the hospital and the physicians’ organization. She chairs MGH’s Quality and Safety Measurement Steering Committee, and one of the committee’s goals is to encourage clinical experts at the institution to participate in national measurement and reporting exercises, to comment actively and provide active feedback to organizations like NQF, CMS, and The Joint Commission. She has in-depth familiarity with key national data sets used at MGH such as NSQIP, Vermont Oxford, The Society for Thoracic Surgery data set, and the American College of Cardiology data set.

Dr. Charles Moseley, Ed.D.
Associate Executive Director- National Association of State Directors of Developmental Disabilities Services

Dr. Moseley became the Director of the Division of Developmental Services for the State of Vermont in 1988. He worked with the NASDDDS staff the directors of six other states and the Association's partners at the Human Services Research Institute to develop and operationalize the National Core Indicators (NCI) developmental disabilities performance assessment system to provide state developmental disabilities agencies with valid and reliable statistical tools to track service outcomes and performance trends, perform state to state comparisons and improve service delivery over time. He is currently working with five states, including Connecticut, Michigan, Maryland, South Carolina and Virginia to assist them in developing the capacity to gather and report NCI outcome and performance data.

Dr. Eugene C. Nelson, DSc, MPH
Director, Population Health Measurement Program, The Dartmouth Institute; Director, Population Health and Measurement, Dartmouth-Hitchcock Medical Center; Professor, Community & Family Medicine and of The Dartmouth Institute for Health Policy and Clinical Practice
Dartmouth Medical School

Dr. Nelson is a national leader in health care improvement and the development and application of measures of quality, system performance, health outcomes, value, and patient and customer perceptions. In the early 1990s, Dr. Nelson and his colleagues at Dartmouth began developing clinical microsystem thinking. His work developing the “clinical value compass” and “whole system measures” to assess health care system performance has made him a well-recognized quality and value measurement expert. He is the recipient of The Joint Commission’s Ernest A. Codman award for his work on outcomes measurement in health care. He received an AB from Dartmouth College, a MPH from Yale University and a DSc from Harvard University.

Dr. Kenneth Ottenbacher, PhD, OTR
Russell Shearn Moody Distinguished Chair in Rehabilitation, University of Texas Medical Branch at Galveston

Kenneth J. Ottenbacher holds the Russell Shearn Moody Distinguished Chair in Rehabilitation at the University of Texas Medical Branch in Galveston. He serves as Senior Associate Dean and Director of the Division of Rehabilitation Sciences in the School of Health Professions. He is also Associate Director for the Sealy Center on Aging. Dr. Ottenbacher received his PhD from the University of Missouri-Columbia and is a licensed occupational therapist. His research interests include rehabilitation outcomes with a focus on functional assessment, disability and frailty in older adults. Dr. Ottenbacher’s research has been supported by continuous federal funding since 1986. He is a member of several editorial boards.

Dr. Greg Pawlson, MD, MPH
Executive Director, Quality Innovations- BlueCross BlueShield Association Office of Policy and Representation

Prior to BCBSA, Dr. Pawlson was Executive Vice President of the National Committee for Quality Assurance (NCQA), where he was responsible for oversight of all activities related to research and analysis, development of performance measurement measures related to most major diseases, and more recently measures related to overuse, resource use and cost. He also served as NCQA's primary liaison to physician specialty societies and medical boards including the American College of Physicians, the American Academy of Family Physicians and the American Board of Internal Medicine. Before NCQA, Dr. Pawlson served as Senior Associate VP for Health Affairs and Medical Director for Quality and Utilization Management for the faculty practice at The George Washington University Medical Center.

Dr. Eleanor M. Perfetto, PhD
Senior Director- Pfizer

Dr. Perfetto holds BS and MS degrees in pharmacy from the University of Rhode Island, and a PhD from the University of North Carolina School of Public Health. She currently serves as a board member of the Pharmacy Quality Alliance (PQA), and is co-chair of the PQA Research Coordinating Council. Prior to joining Pfizer, Dr. Perfetto provided research consulting services for over eight years to government agencies, the pharmaceutical industry, and professional organizations. Prior to consulting, she established a new division responsible for global health outcome and economic research at Wyeth-Ayerst. She also served in the U.S. Public Health Service as senior pharmacoepidemiologist within the Agency for Health Care Policy & Research (now AHRQ).

Ms. Collette M. Pitzen, BSN, RN, CPHQ
Manager, Measure & Program Development- Minnesota Community Measurement

Collette Pitzen is the manager for Measure and Program Development at MN Community Measurement, a non-profit organization whose mission is to accelerate the improvement of health by publicly reporting health care information. She has 27 years' experience in a variety of health care settings including neurology, cardiovascular and dialysis with a significant portion devoted to quality improvement, measure design and reporting. Prior to her current position, Collette worked for Fairview Physician Associates and was responsible for implementing a Clinical Excellence and internal reward program for FPA's primary and specialty care clinics. Collette holds a BS degree from the University Of Minnesota School Of Nursing and is a Certified Professional in Healthcare Quality.

Ms. Cheryl Powell, MS
Deputy Director, Federal Coordinated Health Care Office- Centers for Medicare and Medicaid Services

Cheryl Powell is the Deputy Director of the Medicare-Medicaid Coordination Office (Federal Coordinated Health Care Office) at the Centers for Medicare & Medicaid Services. As the Deputy Director, she assists the Director in leading the work of this office charged with more effectively integrating benefits to create seamless care for individuals eligible for both Medicare and Medicaid and improving coordination between the federal government and states for such dual eligible beneficiaries. She has extensive experience in both Medicare and Medicaid policy development and operations. She is an expert on Medicaid reform activities and policy development. She earned a master's degree in public policy from the John F. Kennedy School of Government at Harvard.

Dr. David Radley, PhD, MPH
Senior Policy Analyst and Project Director, Institute for Healthcare Improvement

David C. Radley, Ph.D., M.P.H., is a Senior Policy Analyst and Project Director at the Institute for Healthcare Improvement. Through a grant from the Commonwealth Fund, Dr. Radley oversees development and production of the Commonwealth Fund's national, state and sub-state health system performance scorecard series. His methodological expertise is in health system performance measurement and in studies that use large administrative and survey-based datasets. Dr. Radley received his Ph.D. from the Dartmouth Institute for Health Policy and Clinical Practice in 2008.

Mr. Ted Rooney, RN, MPH
Project Leader, Maine Quality Counts

Ted is Project Leader for the Maine Health Management Coalition's Pathways to Excellence initiatives, which measure and report the value of health care, and work to change the reimbursement system to reward high value care. Ted is also Project Leader for Aligning Forces for Quality, a Robert Wood Johnson Foundation funded initiative led in Maine by Quality Counts in partnership with the Maine Quality Forum and Maine Health Management Coalition. Ted serves on various Boards and Committees: Maine Health Data Organization Board, AHRQ Healthcare Cost and Utilization Project Steering Committee, Quality Alliance Steering Committee National-Regional Implementation Workgroup, and Healthy Choices for ME advisory committee.

Dr. Debra Saliba, MD, MPH
Senior Natural Scientist, The RAND Corporation

Debra Saliba, MD, MPH is the UCLA Anna and Harry Borun Endowed Chair in Geriatrics and Gerontology and Director of the UCLA/JH Borun Center for Applied Gerontological Research. Dr. Saliba is a practicing physician with the VA GRECC and serves as the Strategic Program Lead for Aging and Long-term Care populations in the VA HSR&D Center of Excellence for the Study of Healthcare Provider Behavior. She is also a senior natural scientist at RAND. Her research in quality of care and vulnerable populations has received awards from the Journal of American Medical Directors Association, VA Health Services Research & Development, and the American Geriatrics Society.

Dr. Marcel Salive, MD, MPH
Health Scientist Administrator, Division of Geriatrics & Clinical Gerontology, National Institutes of Health

Marcel Salive, MD, MPH, joined the Division of Geriatrics and Gerontology, National Institute on Aging, in 2010 and oversees the research portfolio on multi-morbidity treatment and prevention, polypharmacy and comparative effectiveness. Marcel has held leadership positions in the Centers for Medicare & Medicaid Services (CMS), National Heart, Lung and Blood Institute, and Food and Drug Administration. From 2003-2010, he served as Director of the Division of Medical and Surgical Services within the Coverage and Analysis Group of CMS and was responsible for developing and maintaining national coverage decisions for Medicare beneficiaries using a rigorous and open evidence-based process.

Dr. Barbara L. Summers, PhD, RN, FAAN
VP, Nursing Practice and Chief Nursing Officer- University of Texas-MD Anderson Cancer Center

Dr. Barbara Summers is Vice President, Chief Nursing Officer and Head, Division of Nursing at MD Anderson Cancer Center. Dr. Summers has led the creation of new frameworks and models to build an organizational culture that promotes patient-centeredness, healthcare safety and quality, and inter-professional collaboration. Her passion for and commitment to excellence includes sustaining highly reliable, patient-focused systems of care. She serves on the Board of Directors of the Institute for Interactive Patient Care, an organization dedicated to empowering patients and improving health outcomes through direct patient engagement.

Dr. Kalahn A. Taylor-Clark, PhD, MPH
Director, Health Policy- The National Partnership for Women & Families

Dr. Kalahn Taylor-Clark currently serves as the Director of Health Policy at the National Partnership for Women and Families. Her primary responsibilities are in shaping and implementing the policy agenda for the National Partnership's major initiative, the Campaign for Better Care, as well as providing strategic policy support on a range of activities related to delivery system reform, including payment reform, quality measurement, reduction of health disparities, consumer engagement, and promotion of patient-centered care delivery and the effective use of health information technology. Prior to joining NP, she led the Patient-Centeredness and Health Equity Portfolio in the Engelberg Center for Health Care Reform at the Brookings Institution.

Dr. Mary Tinetti, MD
Professor of Medicine and Epidemiology, Yale School of Medicine

Dr. Tinetti is the Gladys Phillips Crofoot Professor of Medicine and Epidemiology at Yale School of Medicine and Chief of the Section of Geriatrics. She was the first investigator to identify that older adults at risk for falling and injury could be identified, that falls and injuries were associated with a range of serious adverse outcomes, and that multifactorial risk reduction strategies were effective and cost-effective. She is involved in efforts to translate these research findings into clinical and public health practice. Her current research focus is on clinical decision-making for older adults in the face of multiple health conditions, particularly trade-offs among health conditions and the harms and benefits of commonly recommended treatments.

Ms. Phyllis Torda, MA
Vice President, Quality Solutions Group- National Committee for Quality Assurance

Phyllis Torda is the Vice President for Strategy and the Quality Solutions Group at NCQA. She leads strategic planning for the company and its consulting arm, which provides services to the federal and state governments. In her 15 years at NCQA, she has led a wide variety of activities related to performance measurement and reporting. She is the principal investigator for NCQA's contract with CMS to develop performance measures for the Medicare population and to evaluate Medicare Special Needs Plans. She also leads the development of measures of inpatient psychiatric care and cancer care for CMS. Ms. Torda has participated in development of the CAHPS surveys since the inception of that AHRQ initiative.

Dr. John Wasson, MD
Emeritus Professor, Dartmouth Medical School

Dr. John Wasson is Emeritus Professor of Community and Family Medicine and Medicine at Dartmouth Medical School. He is Associate Director for the Center for the Aging and is a member of The Dartmouth Institute Patient-reported Measure (and Information) Trust. He represents a research team working at The Dartmouth Institute's Patient-reported Measure (and Information) Trust. This team is committed to collaborative development and testing of patient-reported measures. This team also plans to make publically available the best patient-reported measures for health care. He has participated in the development and reliability testing of several patient-reported measures with a particular focus on functional status and patient experience reporting.

Dr. Robert Weech-Maldonado, PhD
Professor and Chair- University of Alabama at Birmingham

Robert Weech-Maldonado, PhD is Professor and L.R. Jordan Endowed Chair in the Department of Health Services Administration, University of Alabama at Birmingham. Dr. Weech-Maldonado is an organizational researcher who examines the impact of cultural competency strategies in reducing disparities in quality and access to care. He is currently the PI in the Patient Assessments of Cultural Competency (PACC) project, where he is developing and testing patient-centered measures of cultural competency. In another project, he developed and tested the Cultural Competency Assessment Tool for Hospitals (CCATH), an instrument that assesses hospital's adherence to the Cultural and Linguistic Appropriateness Services (CLAS) standards.

Ms. Linda Wilkinson, MBA
Coordinator of Patient and Family Centered Care, Dartmouth Hitchcock Medical Center

Ms. Wilkinson is Coordinator of Patient and Family Centered Care at Dartmouth-Hitchcock on D-H's Value Measurement, Quality and Patient Safety team. She helps design and implement strategies and initiatives to support institution-wide Patient/Family Centered Care practices. She builds liaisons with clinicians, staff, management and communities to assure direct engagement of volunteer patients and family members in policy and process design, from planning to co-production of care. Wilkinson manages the recruitment, training and supervision of Patient/Family Advisors (PFAs) who become working partners with professionals throughout the D-H system to assure the patient perspective of the health care experience is an integral part of D-H's planning process.

NATIONAL QUALITY FORUM

Dr. Albert Wu, MD, MPH
Professor- Johns Hopkins Bloomberg School of Public Health

Dr. Wu is a practicing general internist, Professor of Health Policy and Management, Director of the Center for Health Services and Outcomes Research (CHSOR) at the Johns Hopkins Bloomberg School of Public Health, and Director of the DEcIDE center for comparative effectiveness research. His research and teaching focus on patient reported outcomes and quality of care. He was the first to measure the quality of life impact of antiretroviral therapy in HIV clinical trials, and has developed and tested many widely-used PRO measures. He has applied PROs as performance measures for the care of asthma and other chronic conditions. He was President of the International Society for Quality of Life and has authored over 300 peer review publications.

PANEL MEMBER BIOGRAPHIES:
LESSONS LEARNED FROM THE FIELD

STEFAN H. LARSSON MD, PhD
Senior Partner and Managing Director, Stockholm

Stefan Larsson, joined the Stockholm office of The Boston Consulting Group in 1996. He is the leader of BCG's Global Health care Payer and Provider Practice. Stefan is a BCG Fellow with global responsibility for BCG's work in Value based health care. Stefan's Relevant project experience includes:

- Strategy, organizational redesign and operational effectiveness for leading Nordic University Hospitals as well as private Health care provider organizations. Similar assignments for broad range of public and private organizations within and outside Health care
- Has lead BCG's development of national strategy for Swedish health care with a focus on the importance of outcomes registries for health care improvements
- Stefan leads BCG's work to support the development of the New Karolinska Hospital in Stockholm, Europe's largest Public Private Partnership project
- Broad Pharma experience; Numerous strategy assignments, Global Sales Force effectiveness, Manufacturing, Inlicensing strategy; Commercial compliance etc. Has lead major transformation projects in Clinical Development.
- Broad range of Industry Strategy assignments for Biotech and MedTech companies across Europe as well as Health care market strategy for global leaders in telecom, steel and software
- Over 50 Commercial Due Diligence projects for leading investors seeking opportunities in Health Care: Private providers, Pharma, Med.Tech. Biotech.

Academic Background and professional experience prior to joining BCG

- M.D. Karolinska Institute, Stockholm (KI)
- PhD-Studies in Pediatric Nephrology at KI and Harvard Medical School. PostDoc at MRC Human Genomics unit in Edinburgh and EMBL in Heidelberg.
- Associate Professor at the Karolinska Institute, 23 publications in peer reviewed intl. journals

DAVID NUTTAL
Deputy Director for PROMs Programme,
Department of Health, England

David is currently Deputy Director responsible for the Patient Reported Outcome Measures (PROMs) programme at the Department of Health in England where he has worked since 2001. An economist by background, David has worked in a range of policy areas including Patient Choice policy, Day surgery and Social Care resource allocation. During his time with the Department of Health, David has been closely involved with work on PROMs, including the programme of research commissioned from the London School of Hygiene and Tropical Medicine to identify and pilot PROMs measures for key elective procedures. The research programme, which was undertaken between 2004 and 2007, now represents a significant part of the evidence base underpinning the implementation of PROMs from April 2009. David has led the implementation of PROMs across the NHS in England which is now the largest comprehensive collection of patient outcomes data of its type.

ELIZABETH GOLDSTEIN, Ph.D.
Director of the Division of Consumer Assessment and Plan Performance,

Liz Goldstein is the Director of the Division of Consumer Assessment and Plan Performance at the Centers for Medicare & Medicaid Services (CMS). Since 1997 she has been working on the development and implementation of CAHPS (Consumer Assessment of Healthcare Providers and Systems) Surveys in a variety of settings. She is responsible for the Medicare CAHPS surveys, the Part C plan ratings, the star ratings for Medicare Advantage quality bonus payments, Medicare HEDIS data collection, Part D enrollment analyses, and consumer testing for CMS quality tools.

In addition to her work at CMS, she has conducted research and has published articles related to long-term care, home health care, comparative behavior of for-profit and nonprofit organizations, integrated health care delivery systems, child day care, and substance abuse treatment programs.

She received her Ph.D. in economics from the University of Wisconsin in Madison and her B.A. from Wellesley College.

NATIONAL QUALITY FORUM

Patient-Reported Outcomes Workshop
September 11-12, 2012
Workshop Participants (On-Site)

Blum, Steven
Forest Research Institute

Byer, Michael
M3Information

Christensen, Keri
American Medical Association

one Swartz, Lisa
Press Ganey

Dailey, Maureen
American Nurses Association

Dang-Vu, Christine
Brookings Institution

Giovannetti, Erin
National Committee for Quality Assurance

Haenlein, Kelly
Genentech

Henningfeld, Elizabeth
Center for Medicare and Medicaid Services

James III, Thomas
Humana

Jones, Stacie
American College of Emergency Physicians

Keller, Susan
American Institutes for Research

Kennedy, Cille
DHHS/ASPE

Kurth, Kelsey
American Academy of Ophthalmology

Lohnes, Maggie
MITRE Corporation

Mastanduno, Melanie
The Dartmouth Institute for Health Policy & Clinical Practice

Miller, Deborah
Cleveland Clinic

Nelson, Rachel
US Department of Health and Human Services, Office of the National Coordinator for Health IT

Okun, Sally
PatientsLikeMe

Potter, D.E.B.
AHRQ

Rice, Marty
HRSA

Ross, E. Clarke
American Association on Health and Disability

Rubin, Koryn
American Association of Neurological Surgeons

Smith, Heather
American Physical Therapy Assoc

Sun, Denise
MITRE Corporation

Thoumi, Andrea
The Brookings Institution

Vallow, Susan
GlaxoSmithKline

NATIONAL QUALITY FORUM

Patient-Reported Outcomes Workshop
September 11-12, 2012
Workshop Participants (Off-Site)

Baranowski, Rebecca
American Board of Internal Medicine

Bershadsky, Julie
HSRI

Boggs, Julie
GlaxoSmithKline

Butt, Zeeshan
Northwestern University Feinberg School of Medicine

Crawford, Amaris
American Medical Association

Edwards, Todd
University of Washington

Gjorvad, Gina
American Academy of Neurology

Harder, Joel
SCAI

Heinemann, Allen
RIC

Hunt, Gail
National Alliance for Caregiving

Jadczak, Deborah
The Dartmouth Institute of Health Policy & Clinical Practice

Kidin, Lisa
UT MD Anderson Cancer Center

Lavallee, Danielle
University of Washington

Lentz, Lisa
CMS

Mcgonigal, Lisa
Kidney Care Partners

Miller, Lesley-Ann
GSK

Neumann, Holly
Rehabilitation Institute of Chicago

Schwalenstocker, Ellen
Children's Hospital Association

Shahriary, Melanie
American College of Cardiology

Shaughnessy, Linda
Massachusetts Health Quality Partners

Spinks, Tracy
The University of Texas MD Anderson Cancer Center

Swain-Eng, Rebecca
American Academy of Neurology

Tavernier, Susan
University of Utah

Weng, Weifeng
ABIM

West, Michael
GSK