Patient-Reported Outcomes in Performance Measurement

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

October 22, 2012

Prepared for NQF

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

[1]RTI International, [2]Brookings Institution

Introduction

National Quality Forum Initiative

The National Quality Forum (NQF) is embarking on an effort to examine the methodological issues in developing scientifically rigorous performance measures based on patient-reported outcomes (PROs). NQF's project is intended to inform its process for evaluating PRO-based performance measures (PRO-PMs) that can be used in quality-improvement and accountability activities, including public reporting and performance-based payment.

To support this initiative, NQF commissioned two papers on the use of PROs in performance measurement. The first paper reviewed the instrument-level issues that should be considered when selecting PRO instruments or measures (PROMs), for use in performance measures.[1] This paper, the second in the series, is intended to help inform next steps on the path to developing PRO-PMs by examining methodological issues specific to using PROMs for measuring the performance of healthcare organizations.

The NQF has undertaken extensive work in the area of outcome performance measures over the last few years. The *National Voluntary Consensus Standards for Patient Outcomes: A Consensus Report* defined outcomes as being important because they "reflect the reason that an individual seeks healthcare services."[2] The individual patient's voice in many performance measures, however, has largely been missing. Few PRO-PMs are available at the organizational level, even though patients are often the best able to report on the experiences and results of their individual care.[2] Both commissioned papers address these problems; this paper goes beyond measurement of individual patient outcomes to consider development of reliable and valid performance measures that various provider groups, regulatory agencies, payers and insurers, and others can use for accountability and quality-improvement activities.

Purpose of This Background Paper

With a focus on measures of provider performance, this paper examines some key methodological issues and concerns. Several of these apply to item or instrument development (e.g., issues of reliability and validity), as discussed in the first paper.[1] When aggregating PROM responses across all cases for a provider, additional considerations of validity and reliability associated with how the measure is constructed and interpreted are needed. These include factors such as selecting the PRO items or instruments, defining the appropriate target population, and establishing exclusion and inclusion criteria that ensure adequate samples can be found in the individual provider organization. Further, the approaches for calculating performance scores and using risk-adjustment techniques to adjust for differences in population case mix must also be considered.

The selection of priorities for performance measurement and identification of performance gaps is work that the NQF conducts through its National Priority Partnership, and is not addressed in this paper. The importance of a performance measure, defined as the extent to which patient outcomes may be improved, is a key evaluation criterion for all performance measures.

Important Terminology

The crucial point, for this paper, is that NQF reviews and endorses performance measures – in this context, PRO-PMs. NQF does not endorse either PROs per se (e.g., fatigue, pain, depression, functional status) or PROM. Given the centrality of these distinctions, we define key terms used in this paper as they are used by NQF in this initiative:

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.

PRO patient-level measure/instrument (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 a 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).

Provider: A clinician, facility, or organization.

Reliability: 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).[3]

Validity: 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).[3]

Key Components of a Performance Measure

An organizational performance measure is a numeric quantification of healthcare quality for a designated accountable healthcare entity (i.e., provider), such as a hospital, health plan, nursing home, or clinician. When evaluating a performance measure, carefully considering its key components is critical. They include the following four elements: (1) the item or instrument that measures the health concept of interest; (2) the calculation of the performance score; (3) the target population and inclusion and exclusion criteria; and (4) the risk adjustment methodology.

For a PRO-PM, the selection of the items or the instrument that measures the health concept at the individual level—the PROM—is central to the scientific acceptability of the eventual PRO-PM. In the first white paper, Cella et al. identified the following eight important characteristics to consider when evaluating and selecting PROMs:[1]

1. conceptual and measurement model (defining and describing the concept[s], how the concept[s] are organized into a measurement model),
2. reliability (internal consistency for multi-item scales, reproducibility),
3. validity (content, construct, criterion, responsiveness),
4. interpretability of the scores (reference population normative values, guidance on the minimally important difference in scores that can be considered meaningful from the patient or clinician perspective),
5. burden (time, effort, and other demands on the respondent and administrator),
6. alternative sources, modes, and methods of administration,
7. cultural and language adaptations, and
8. electronic health records (interoperability, electronic data capture).

Each of the eight PROM characteristics listed above should be considered in evaluating the scientific acceptability of a PRO-PM; however, a PRO-PM should not be excluded from consideration if it meets many, but not all, of the characteristics. Building a PRO-PM across all relevant patients in an organization may be more complicated than assessing outcomes at the individual-patient level. For example, pain and depression symptoms are important health issues to measure as a PRO, but collection of these data from all patients across an entire organization may be challenging. For patients who are unable to self-report their pain status or depressive symptoms, data collected from a proxy based on patient observation may not be equivalent to the patients' self-report. Therefore, requiring the use of alternative sources for PROM data collection may not always be appropriate within a PROM-PM. Similarly, experience with care is another important PRO concept. However, patients' perceptions or reports about a clinician or clinical team should not be incorporated into the patient's electronic health records

(EHRs), because these data are not intended to be shared with providers at the individual patient level.

Five of the eight characteristics address critical issues that should be required evidence in a PRO-PM:

- conceptual and measurement model
- reliability in the target population
- validity for the target population
- interpretability of scores
- burden

Four of these five characteristics map to NQF's evaluation criteria of importance (conceptual and measurement model), scientific acceptability (reliability, validity) and feasibility (burden). The fifth characteristic, interpretability of scores, refers to the availability of reference population normative data and guidance on the minimally important difference in scores that can be considered meaningful from the patient or clinician perspective. Interpretability of scores may be a key issue to consider when patient-level data are aggregated to the provider-level performance score of the PROM-PM. When evidence supporting the three remaining 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)—is available, the PRO-PM may be more robust than alternatives under consideration. However, these three characteristics should not be required for all PROM-PMs.

A second component of a PRO-PM is the calculation of the performance score. This is a provider-level value (derived from aggregating individual-level PROM data) that is meant to evaluate quality and distinguish levels of quality of care or, at least, good and poor quality. The PROM data used in performance measures may be the score from an individual item (e.g., a pain score) or a scale score derived from an instrument or set of items (e.g., a PHQ-9 score). Performance scores can take many forms; the main forms are a count, a percentage, a mean, a median, or a ratio value. We describe these types of scores and give some examples below in the section describing performance-measure validity considerations.

A third component of a performance measure is the target population, meaning the individuals to be included in the performance measure. This element involves specifying both inclusion and exclusion criteria. The target population for a PRO-based performance measure should be selected based on literature documenting the testing of the PROM in the target population. Some PROMs are classified as generic and could apply to a wide range of individuals, including those with multiple chronic conditions, while other PROMs are condition-specific and are only appropriate for use with a limited target population.[4] For some populations, collection of both generic and disease-specific measures may be valuable so that both specific and broad health status data may be examined. Condition-specific measures provide information that is directly relevant to the patients' condition, but they cannot be compared across conditions. Generic measures facilitate the comparison of outcomes across conditions.

Risk adjustment is a fourth key component of an outcome-performance measure. These techniques (discussed in more detail below) adjust for patient case-mix differences across providers or across time, so that the performance of the provider can be determined. Most, but not all, outcome measures need to be risk adjusted so that observed differences to the providers' care reflect provider performance and not population differences.[5,6]

All these key components—the selected PROM, the performance score derived from the PROM data, the target population and exclusion criteria, and the risk adjustment methods—contribute to the ability of a PRO-PM to yield a valid estimate of the quality of care provided. Given the importance of the performance score, and the focus of this paper, the next section of this paper describes options for calculating this score.

Defining the Performance Score Using Aggregated Data

There are several options for aggregating patient-level PROM data into a provider-level PRO-PM. The approach to creating a discriminating performance measure will vary depending on the PROM and the distribution of the concept it measures in the patient population. The outcome measure (i.e., endpoint) for PROMs will be classified as either a desirable or undesirable patient-reported outcome or health status.[7] For both desirable and undesirable outcomes, there are several options for reporting the provider-level outcome or health status including: (1) a change in status (e.g., amount of decrease in pain symptoms between start of care and end of care, decrease in pain impact on sleep, amount of increase in functional status between start and end of care, no change in functional status if avoiding functional decline is a goal)[7,8] or (2) a threshold achieved (e.g., percent of patients with moderate to severe pain, pain no longer impacting sleep).[8]

While the change in health status, calculated as the difference between the follow-up score and baseline score, such as reduction in pain or improvement in functional status, may initially seem like the best choice for measuring the effectiveness of care, there are several methodological limitations with calculating change scores.[9] One problem with calculating a mean change score at the provider level is that each individual's change score 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 both the baseline scores and 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 may not be sensitive enough at the low end or high end of the scale to detect the change. Finally, for many health status measures, the clinical meaning of a change score is unknown, and the change score is hard to interpret.[10]

Some health status measures have thresholds that define "stages" or "complexity" levels that are clinically meaningful,[11-16] and moving from one clinically meaningful stage to another stage may have meaning to the patient. A minimal detectible change or reliable change index score can be calculated for health status measures, but these values provide an estimate of the amount of change that would reflect more than measurement error;[9] they are not meant to document a change that is meaningful to the patient or clinician. For PROMs with established cut scores, such as the PHQ-9 scores (> 9 indicating depressive symptoms), the cut points were developed to screen for a diagnosis or the need for treatment, not to assess meaningful changes in patient status. The PRO-based performance measure "Change in basic mobility as measured by the AM-PAC" (NQF #0429) uses a change score to document improvement in functional status, while the PRO-based performance measure "Percent of residents with moderate to severe pain (short stay)" (NQF # 0676) uses a point-in-time threshold value. The performance measure "Depression Remission within 6 Months" (NQF # 0711) uses a threshold value, but it 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 for selecting a change or threshold depends on the measurement goal.

The performance score is calculated at the provider-level and may be a mean, percent, or ratio value. It is meant to measure quality and should 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 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. For example, a mean value may be misleading in situations where the population is heterogeneous because patients' change score or threshold score will vary a great deal, and no single value would represent the diversity of the patient population.[17]

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 based on a PROM-specific clinically important difference; or (3) a threshold value that is associated with a longer-term outcome (e.g., a patient's fear-of-falling score, which is associated with a reduced risk of falls). For some PROMs, improvement may be reflected in higher scores, while for other PROMs, lower may indicate improvement. For PROMs that have established clinically meaningful thresholds (i.e., cut points), the performance score could 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 > 9 has been found to be clinically important, with a sensitivity of 88% and a specificity of 88% for a clinical diagnosis of major depression.[18] Sensitivity refers to how often the test will be positive (true positive rate), if a person has a condition. 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," (NQF # 0711) 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 to 27 = moderate to severe depressive symptoms), and the patient is considered to have made an improvement if he or she moves from the moderate to severe depressive symptoms category (> 9) to the not depressed category (< 5). The performance measure "Change in basic mobility as measured by the AM-PAC" (NQF # 0429) 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. The minimal detectible change can be calculated using the standard error of measurement (SEM) and a selected confidence interval (e.g., 90%, 95% or 99%). The formula for the 95 percent confidence interval minimal detectible change = $1.96 \times \text{SEM} \times \sqrt{2}$.

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 varies and adjustment for the time exposure is appropriate.

Each of these approaches has been used in different measurement efforts. Selecting the right approach depends on the type of PROM being examined. Selection of the method depends on the type of question being answered and whether the proposed approach provides a robust measure given the constraints of the measurement design.

In the next section, we build on the NQF's Measure Testing Task Force Report[3] and describe issues related to the reliability and validity of the performance measures. We also identify unique issues that are specific to performance measures based on PROMs rather than clinician assessment of outcomes or process descriptions.

Key Issues of Reliability and Validity For Performance Measures

Reliability

Reliability is an important measurement concept as it refers to the repeatability or precision of measurement.[3] Reliability can also be conceptualized as consistency 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 attributable to systematic differences across the measured entities (or signal) in relation to random error (or noise).[3] Data-element reliability does not guarantee reliability of the provider-level performance measure (clinician, facility, or organization). Also, reliability is a necessary, though not sufficient, pre-condition for validity; this is true for both PROMs and PRO-PMs.

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.[19,20] Below we expand on the definition of reliability given above and describe methods for evaluating the reliability of PRO-PMs for all types of providers. We also offer strategies for improving reliability when test results indicate that measure reliability may be a concern. While we focus on provider-level measures in the discussion below, many of the principles and methods discussed are also applicable to the evaluation and design of population-level measures.

Defining Patient-Level Reliability and Provider-Level Reliability

In their classic text, *Health Measurement Scales*, Streiner and Norman supply a general conceptualization of reliability that can be applied at either the patient or provider level and can be used when thinking about the reliability of item, instrument, scale, or performance measure.[9] They describe reliability as the ratio of the true variability among subjects over the total variability among subjects. This can also be stated as the following formula, which is a form of the intraclass correlation coefficient:

reliability = subject variability/(subject variability + measurement error)

This reliability coefficient decomposes the total variability among subjects into two parts: (1) true between-subject variability and (2) measurement error.[9] When evaluating the reliability of a PROM, the subjects in the above equation are patients; when evaluating the reliability of a PRO-PM, the subjects are providers. Therefore, applying this general formula at the level of individuals (person, patient) describes the reliability of the PROM as the ratio of the variability among patients' PROM scores to the total variability among patients PROM scores.[9]

Reliability for the PRO-PMs is conceptually the same. It can also be represented by the following formula:

reliability = signal/(signal + noise)

Adams describes this as "the squared correlation between an observed measurement and the true value of the measure."[19] *Signal* can be interpreted as the true difference between providers' PRO-PM scores. *Noise* can be interpreted as the measurement error incorporated into that PRO-PM. Noise may be introduced by patient-level variability, which may include unmeasured patient characteristics not incorporated into risk-adjustment and lack of reliability of the underlying PROM at the individual level. Noise can also be introduced by lack of precision, or variability, in the PRO-PM calculation due to lack of sufficient patient sample size among providers being evaluated for that PRO-PM.

This reliability coefficient 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 attributable to true differences among providers.[19,21] The reliability coefficient provides information about the extent to which the variability in observed values of the PRO-PM is attributable to true differences in quality, assuming that the PRO-PM has been validly specified, including appropriate adjustments for differences in patient characteristics that might impact provider PRO-PM scores independent of the quality of the provider. Thresholds cited in the literature for reliability, estimated using signal-to-noise calculation, for performance comparisons among groups of providers is 0.70, and among individual providers (e.g., when a patient is selecting a provider based on performance-score information) is 0.90.[8] In the next section, we describe methods for statistical evaluation of the reliability of PRO-PMs. Some methods provide reliability estimates for each provider rather than a single summary reliability estimate for a PRO-PM.

Provider-level reliability is determined by three factors: (1) the magnitude of true differences among providers, (2) the within-provider variation, and (3) the size of the provider sample (i.e., denominator).[19] Reliability is therefore not an intrinsic characteristic of a measure. Rather it depends on the characteristics of the set of providers and patients included in the measure specifications. For example, regarding the magnitude of true differences among providers, if most providers perform well on a particular PRO-PM and are therefore clustered together (ceiling effect), the PRO-PM may suffer from low reliability. This is one way in which the approach to aggregating PROM data to the PRO-PM may impact PRO-PM reliability. If the selected threshold or change in PROM is easily reached by most patients (or rarely reached), the result may be reduced true variation among providers and therefore reduced reliability of the PRO-PM. However, if patient-level variation is not well accounted for by risk-adjustment (within-provider variation), PRO-PM reliability may be threatened because observed differences among provider PRO-PMs may actually be largely attributable to differences in the providers' patient populations rather than true differences in provider quality. Lastly, estimates for providers with fewer patients or low case volumes are more vulnerable to random error than are estimates for larger providers and therefore likely to have less stable or less reliable PRO-PM scores. Therefore, reliability testing results cited during the NQF measure submission process should be specific to the population that the PRO-PM is designed to capture. We can also conclude from these determinants of reliability, that the reliability of a PRO-PM is not static. For example, if all providers improve their true performance on the PRO-PM, thus reducing the true variability among provider performance measure scores, the reliability estimate will decrease if the sources of noise remain static over the same period.[19]

Testing Reliability

There are multiple methods for testing reliability of a provider-level performance measure (e.g., signal to noise). In this section, we summarize how the following methods can be used to evaluate PRO-PM reliability: hierarchical modeling, confidence-interval estimation, inter-unit (F-test) reliability testing, intraclass correlation, generalizability theory (factorial analysis ANOVA), and Monte Carlo simulation.

Two-level hierarchical models can be used to estimate signal and noise. Researchers at the RAND Corporation have published a tutorial for this approach.[19] These models control for random error at the patient-level and at the larger, organization level and allow estimation of signal and noise that can be used to calculate the reliability coefficient described in the opening of this section. The specific modeling approach varies depending on the specification of the performance measure. For instance, 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.[19, 22] Hierarchical linear models are appropriate for composites or continuous measures. Regardless of model type, using hierarchical modeling yields a reliability estimate for every provider. These provider-specific estimates can be used to evaluate the reliability of specific provider PRO-PM scores and to evaluate the overall reliability of the measure. Thresholds cited in the literature for reliability for performance comparisons among groups is 0.70, and among individuals is 0.90.[19] If the reliability indexes for a large proportion of providers fall below a threshold of 0.70, reliability may be a concern for the performance measure.

Other strategies for estimating the random error around a performance measure include reporting provider-level performance measures with an estimate of uncertainty, such as a confidence interval for each provider's performance measures. This method has some drawbacks as it does not provide a summary reliability score to evaluate the PRO-PM nor a provider-level reliability coefficient to evaluate the reliability of a particular provider's score. In addition, this method may be impractical when evaluating very large numbers of providers simultaneously. However, by examining the overlap of confidence intervals, this approach provides a means for visually examining the extent of random error relative to the differences in provider-level scores.[23, 24] If the confidence intervals for individual providers' PRO-PM scores across the range of PRO-PM performance show high overlap compared to the distribution of providers' point estimates, the PRO-PM has low reliability. Similar to hierarchical modeling, the specification of the performance measure determines the appropriate statistical approach for calculating the confidence interval, with different formulas available to calculate confidence intervals for binary outcomes and for continuous measures.[23]

Another systematic strategy for estimating reliability is to calculate what Zaslavsky[23] calls the interunit (here, we would call it inter-provider) reliability. This is the proportion of the variance in a measure that can be attributed to true differences in provider performance; it can be calculated using the value of F derived from an F-test by using statistical procedures such as SAS PROC GLM.[23] Interunit reliability can be specified as 1- (1/F) and can be interpreted similarly to the reliability score described in the introduction of this section, where 0 represents a PRO-PM with variability that is entirely attributable to error, and 1 represents a PRO-PM with variability that is entirely attributable to true differences among provider performance. Unlike the hierarchical models described above, however, this interunit reliability calculation does not yield a reliability estimate for each provider. Rather, it is a summary metric that can be calculated when the standard errors of estimates for providers are similar.[23]

The intra-class correlation coefficient has also been used to calculate measure reliability.[25] Hofer et al. describes calculating reliability based on the mean of patient values within provider and the intra-class correlation coefficient using the Spearman-Brown prophecy formula.[26] The interpretation is similar to that for the other reliability indexes described above:

the provider (or subject) variance over the total variance, which includes the provider variance and variance attributable to measurement or random error. For example, given a calculated reliability of 0.70 for a given measure, one would interpret that value as indicating that 30 percent of variation in that measure arises from chance.[26] The intraclass correlation coefficient also yields a reliability coefficient estimate for each provider, similar to the hierarchical modeling approach.

Another method for evaluating reliability includes generalizability analysis based on generalizability theory on sources of variation, which uses a factorial analysis of variance (ANOVA).[27,28] This method allows for the calculation of components of variance including sources of error. This method is dependent on the ability of evaluators to measure potential sources of variance and error (e.g., clinician, surveyor vendor, day of assessment, patient). Factorial analysis ANOVA yields an estimate of variance attributable to each measured source of variance. This approach also includes an interaction component, which can take into account variation between sets of patients and providers, error variance that is attributable to the design of a measure, and residual variance, which is error from sources that are unmeasured.[27,28] Therefore, results can be used to quantify the reliability of a PRO-PM using the proportion of variation in a measure that is attributable to error using the signal-to-noise equation described in the opening of this section. Because generalizability theory can be used to quantify sources of variance and error in a PRO-PM, this approach can also potentially provide targets for improving reliability by decomposing error into its potential sources. For example, for a survey-based PRO-PM collected by different survey vendors, generalizability analysis could be used to identify whether error is increased for providers using one survey vendor compared to providers using a different survey vendor.

The last method for evaluating reliability that we describe is Monte Carlo simulation. The Monte Carlo simulation method randomly simulates patient data for providers being evaluated, taking into account observed between-patient and between-provider variability in patient characteristics. Investigators also simulate patient outcomes and calculate performance measures based on multiple iterations of simulated data to evaluate performance measure reliability. For example, in a study that examined the reliability of four Bayesian measures of hospital mortality, investigators estimate the reliability of the measures by examining the correlation between pairs of the measures based on 100 Monte Carlo simulations.[29]

Addressing Sample Size Issues

Two important factors to consider in designing performance measures are: (1) the sample size needed to provide a reliable estimate and (2) the likelihood that providers will achieve an adequate sample size for a specific level of analysis (e.g., clinician or hospital). Evaluators should consider the minimum sample sizes needed to calculate PRO-PM scores reliably for individual clinicians and other levels of provider organizations; they also need to understand how many providers would fall below this minimum sample size.

General sample size cutoffs commonly used to identify provider results most vulnerable to random error for exclusion from public reporting may be insufficient because reliability does not depend on sample size alone, but, as described above, depends on the real differences between providers and within-provider variation.[19] Signal-to-noise calculation of reliability can allow some estimation, specific to a particular PRO-PM, of a minimum sample size necessary to meet the usual reliability thresholds (0.70 for group comparisons and 0.90 for individual comparisons). Examining the relationship between individual providers' reliability estimates and their measure denominator sizes can be useful for identifying the threshold denominator size where reliability index estimates for a PRO-PM are 0.70 or greater.[19,22] Prior studies have graphed individual providers' performance-measure-denominator counts against their reliability coefficient scores to determine this threshold. [19,22]

Methods for Improving PRO-PM Reliability

If reliability estimates show that a PRO-PM has poor or marginal reliability (i.e., the reliability indexes described above fall below 0.70), or if large proportions of providers have reliability estimates below this threshold, evaluators may need 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 the number of data points being used and increasing the stability of the performance measure.[25] However, PRO-PMs will need to meet NQF standards and recommendations for valid composite design, and it will be important to consider the conceptual appropriateness of combining different individual PRO-PMs.[30]

Another approach 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 in the measure denominator. A potential drawback to this strategy is that the measure may be insensitive to changes in quality over time. Thus, it may be unacceptable to providers that want to make appropriate changes in the quality of the care they deliver.

The reliability of the PRO-PM can also be enhanced by improving the reliability of the underlying PROM. Improving the reliability of the PROM at the individual level can reduce the contribution of noise to the overall variability in the observed PRO-PM. The first paper[1] includes an extensive discussion of potential sources of variability that could be considered for targeting at the item or scale level to improve PROM reliability and therefore the reliability of any PRO-

PM based on that PROM. Reliability testing results cited during the measure submission process should clearly identify the version of the instrument used during testing if revisions have been made to items to improve the reliability of the underlying PROM.

Another strategy for reducing random error in a performance measure is to apply a reliability adjustment. This method shrinks PRO-PM estimates toward the mean value for providers in proportion to the amount of random error in the PRO-PM estimate for each provider. Therefore for smaller providers, which are more vulnerable to random error, shrinkage toward the mean is greater. By shrinking PRO-PM estimates toward the mean, the random error in the PRO-PM estimate is reduced and reliability improved.

Estimating performance measures using hierarchical modeling with empirical Bayes shrinkage estimators is one strategy for accomplishing this.[21,23,31,32] One potential concern about this strategy, however, is that the PRO-PM score resulting from the shrinkage adjustment, called here the "shrinkage estimate," may mask a poor provider's performance. If the true performance of a provider is quite poor compared with the mean, the shrinkage estimate will mask that difference by pulling the provider performance measure score toward the mean. Additionally, smaller providers that have values that indicate higher quality than the mean will be pulled closer towards the mean.[21] An alternate strategy to potentially handle these criticisms is to shrink estimated provider performance measures towards the mean value expected for a provider of similar size (e.g., annual volume of ambulatory visits, annual volume of inpatient admissions).[21] However, given the evidence connecting higher volume to quality for some patient outcomes, particularly in acute hospitals, experts debate the desirability of using provider size for reliability adjustment if size is being measured as volume of patients. [32] Volume is a potentially endogenous variable because prior low quality may be the cause of the current low volume. The debate over the use of volume alone is not settled.[21]

One way to address this potential criticism is to shrink the provider performance measure towards the mean value expected for a provider of that size and other attributes (e.g., rural versus urban location).[24] Alternatively, if size can be measured using a metric other than volume per se, such as numbers of beds, the potential endogeneity can be reduced.[24] Examples of this method of using size and other provider attributes for reliability adjustments applied to performance measures are available.[24] Other limitations to the strategy of shrinkage toward the mean are cited in the literature as well; for example, a study examining hospital mortality performance showed that this strategy did better at identifying the "best" hospitals than it did identifying the "worst."[21] However, authors suggested that this pattern may have been attributable to the low mortality rates for the surgical procedures being examined, which resulted in more small hospitals with no deaths than small hospitals with 100 percent deaths, so more small hospitals were moved from the best category than from the worst category when shrinkage estimators were applied.[21]

In summary, it may not be possible to get an accurate PRO-PM estimate for all providers; however, hierarchical modeling with Bayes shrinkage estimators does provide a strategy for improved PRO-PM estimates for providers that previously may have been too small and unstable to yield reportable measures. This method is being used for performance measures on the Centers for Medicare and Medicaid Services (CMS) Hospital Compare website and is recommended by both CMS and the Agency for Healthcare Research and Quality (e.g., NQF #1789: Hospital-Wide All-Cause Unplanned Readmission Measure; NQF #0272: Diabetes short-term complications admission rate).[32]

These approaches to assess the reliability of performance measures apply to PRO-based performance 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. Thus, the approach 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,[3] *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 measure* 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 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 criteria for selecting the target population (denominator), and the risk-adjustment procedures will determine whether a valid and reliable PROM can be a used to create a valid and reliable PRO-PM.

Validity Testing

Strong evidence of validity at the performance-measure level can be challenging to demonstrate, particularly for newly developed performance measures. Validity testing often begins with face validity, which refers to the credibility of the measure based on expert review. Face validity can be tested using a systematic process such as a modified Delphi survey,[3] a formal consensus process, the UCLA/RAND Appropriateness Method,[33] or the American College of Cardiology and American Heart Association Methodology for the Selection and Creation of Performance Measures.[34]

Given that PROMs represent the patient's perspective, face validity of PRO-PMs could also be tested with "patient experts" by 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 complex concept of healthcare quality in a way that these individuals can understand. Hibbard[35] 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 in detail.

Validity of the PRO-based performance measure may also be tested based on criterion validity, which refers to the extent that the measure agrees with a "gold standard." Strictly speaking, there are no gold standards in PRO performance measurement. However, for practical purposes, researchers may use 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).[3] For a PRO-PM, 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-PM may be established. Construct validity refers to how the measure performs based on theory.[3] Construct validity could potentially be tested by comparing known groups, for example, identifying providers that are "centers of excellence" for a PRO construct or the target population and comparing these providers' performance scores with the performance scores of providers that are not considered centers of excellence. Centers of excellence should have performance scores that indicate higher quality. Construct validity might also be tested by comparing the outcomes of patients who are receiving care that is known to differ in the ways that should affect the PROM-PM. For example, during or shortly after early adoption of best practices, providers should have performance scores that indicate higher quality than that of providers who have not adopted those practices. Another example of demonstrating construct validity would be showing a correlation between providers' performance on a process measure with performance on an associated outcome 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 some PROMs addressing depression symptoms, important features are 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" (NQF # 0711) does recognize these time frames and requires collection of follow-up data at 6 months.

Another important issue related to PROMs is that instruments that measure symptoms, such as depression symptoms, are screening tools and are not equivalent to a clinical diagnosis of depression. Therefore, a clear definition of the intended target population (e.g., individuals with symptoms of depression or individuals with clinical depression) and any inclusion or exclusion criteria need to be specified. The performance measure "Depression Remission within 6 Months" (NQF # 0711) handles this issue through its denominator inclusion criteria, which require 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 means that only those patients with the clinical diagnosis and significant symptoms 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 and reduction of pain may be the primary treatment goal for certain populations, such as patients with low back pain, and a PRO-PM 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 the entire patient population cared for by a provider. For example, the performance measure "Percent of Residents with Moderate to Severe Pain (Long Stay)" (NQF # 0677) is targeted for nursing home residents and focuses on the percentage 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 field of pain measurement is moving from using a numerical rating scale that does not take into account individual pain tolerance to measures of the pain's impact on sleep or other activities. (See recent work by the CMS on measuring pain in the CARE item set (http://www.pacdemo.rti.org/). Parallel efforts to document the impact of symptoms related to medical conditions such as headaches and asthma are also underway.

Within the domain of health behaviors, which includes behaviors that are potentially detrimental to health, such as smoking and excessive alcohol intake, the initial prevalence of the behavior may vary by geographic region and thus may vary at the provider level (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; some providers may have very small number of patients who smoke. 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.

A unique issue for the concept of health-related quality of life (HRQL) is that it is a multi-dimensional construct that covers physical, social, and emotional well-being.[36] It may be more challenging to include the HRQL concept within a performance measure in a way that reflects the quality of services furnished by the provider, because factors other than care provided by a healthcare entity can affect an individual's HRQL. Evidence documenting healthcare interventions that lead to improved HRQL in the target population will be particularly important for a HRQL performance measure in order to demonstrate validity as an indicator of quality. Performance measures based on HRQL may best be targeted to homogeneous populations, such as individuals with a knee replacement or spinal cord injury.[37]

Threats to Validity of PRO-based Performance Measures

There are many potential threats to validity for performance measures and PRO-PMs in particular. Threats to the validity of a performance measure can be classified into three broad categories: item or instrument validity, missing data due to non-response or other reasons, and inadequate or no case-mix adjustment when case-mix differences exist. Each of these topics is discussed below.

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 PROM. 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 experience with care measures,[38] and an uneven distribution of these patients across providers 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)" (NQF # 0676) 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 CMS. A script for asking the patient about pain is included on the MDS 3.0 form, and this may support inter-interviewer reliability. For the PRO-based performance measure called "Depression Remission within 6 Months" (NQF # 0711), 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 may not be 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 diagnosed as clinically depressed, are not included in the denominator, and thus, they not included in the calculation of remission at 6 months.

PROMs should be tested to determine if the PROM item(s) have the same meaning across all patients within the target population. Patients from different backgrounds or characteristics (race, ethnicity, gender) who have the same health status may respond differently to items in a systematic way. This is called differential item functioning. For some PROMs, such as the PHQ-9, testing for these subgroup differences has included examining the factor structure of the items across race/ethnicity groups. This research found the factor structure of the PHQ-9 items was similar for Latinos, African Americans, Chinese Americans, and non-Hispanic Whites.[39] Another study, examined the sensitivity and specificity of the cut point of > 9 for the Thai version of the PHQ-9, and found sensitivity was much lower for the Thai translation than populations previously studied with the PHQ-9.[40] Yang[41] found that with the Center for Epidemiological Studies depression scale, women and men responded to the "crying" item differently. If differential item functioning exists among subgroups for a PROM, then the score derived from the PROM could adjust for this effect. If the subgroup differences are not known or are not accounted for in the PROM, the score for the affected subgroups may not be valid.

Testing for differential item functioning within subgroups of the target population is particularly important when computer-adaptive testing is used to collect data. This is because the computer-adaptive test selects a minimal number of items targeted for the respondent. Respondents who do not provide responses in the expected pattern (due to differential item functioning) may be assigned a PROM score that does not reflect their true health status. For example, when the population is heterogeneous, motor functional status might best be separated into the constructs of self-care and mobility rather than one single construct of motor function that combines self-care and mobility. This separation allows the functional outcome scores for the two subscales to vary depending on patients' abilities within each subscale. For example, 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). Patients recovering after a hip replacement have primarily mobility limitations, while patients with central cord syndrome tend to have primarily self-care limitations. Separating the motor scale into two subscales of self-care and mobility skills would result in more precise measurement of function than if data from the patients with both diagnoses were pooled.

An outcome-focused PRO-PM could be specified to use data from more than one PROM if the two (or more) PROMs measure the same construct, such as depression symptoms. However, use of more than one PROM would require research demonstrating that the different PROMs are equivalent as used in the PRO-PM. For example, if the performance measure uses PROM data to assign patients into clinically important groups (e.g., depressed, not depressed), the accuracy of classifying patients into these two groups for both PROMs should be similar. If the research examining the equating process shows the assignment into clinically meaningful groups is not well aligned, the use of more than one PROM 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). Research equated the thresholds, and the performance measure specified the equivalence of these thresholds.[42]

Accounting for Patient Preferences

Another validity issue for PRO-PMs is 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. Calculating a change in pain level or reporting a threshold may not recognize patient preferences related to pain management. An example of a question 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 selected patients from the denominator. However, exclusions need to be supported by evidence.[3] As noted in the CSAC Guidance on Quality Measure Construction,[43] the effect of exclusions for patients preference should be transparent because exclusions for patient preferences (e.g., refusal) may be related to quality-of-care problems.

Missing Data and Response Rates

Another threat to validity occurs when data are missing, but not missing at random. As noted in the CSAC Guidance on Quality Measure Construction,[43] missing data may be indicative of a quality problem itself, therefore, excluding those cases may present an inaccurate representation of quality. Since patients are the primary source 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 or the individuals who do not respond to the survey have different outcomes or experiences than the individuals who do respond, non-response error is a concern.[44]

An additional concern about response rates is that they are often not calculated correctly and are sometimes misrepresented.[4] 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-PM testing. The AAPOR Council has indicated that no single number or measure reflects the quality of a survey and provides the definitions and formulas for calculating response rates, including cooperation rates, refusal rates, and contact rates. The calculations for these four metrics are provided in Appendix A.

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 differing response rates, reporting response rates along with performance scores for PRO-PMs may be important.

Some studies have found that when response rates are low, results are more likely to be biased, either positively or negatively.[44-46] Thus, surveys used for PRO-PMs 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 considerations.[47] In addition, more recent research[47] has focused on the principle of social exchange, which emphasizes that rewards for responding to surveys should outweigh any perceived costs. For example, Dillman[47] 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 each response is important. The respondents’ perceived 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 reviews by one or more expert panels, consumer input, cognitive interviews, and pilot testing. In their analyses of non-responders to the 2007 Medicare Consumer Assessment of Healthcare Provider and Systems (CAHPS) survey, Klein et al.[46] made several recommendations for improving the representation and response to these surveys, including targeted pre-notification materials and campaigns, tailored follow-up, targeted Spanish mailings, Chinese translations/calls, and adjustments to telephone protocols. Some of these recommendations have been implemented, including the development of translations of the Home Health Care CAHPS survey into Spanish, Simplified Chinese, Traditional Chinese, Russian, and Vietnamese.

Missing data can also be a problem when patients cannot respond to a survey due to communication limitations, language barriers, functional limitations, or other reasons. A unique feature of PRO-PMs is the process of data collection, which may need to be flexible in order to accommodate patients' diverse language, cultural, and education backgrounds, as well as diverse functional abilities. Data collection may vary in three 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. 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, some patients may choose not to complete a survey but would be willing to be interviewed. Comparisons of data collected using self-report versus interviewer administration tend to show high reliability; however, this is not always the case.[48] In a clinical setting, interviewers would be clinicians rather than research staff members, and clinicians will have varying skills as interviewers and 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 and inter-interviewer reliability may be appropriate. For the performance measure "Percent of Residents with Moderate to Severe Pain (Short Stay)" (NQF # 0676), 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 is less useful for more subjective PRO concepts, such as pain intensity, nausea, and depression symptoms, because proxy data in this area tend to be less reliable.[49] Proxy responses are reasonable to consider for child health measures where parents are proxies, and the research has shown 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 on No Risk Adjustment

Another potential threat to validity might be differences in case mix with no or inadequate risk adjustment. 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 that are present at the start of care 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.[5,6] Domains to consider for risk adjustment of PRO-PMs and different approaches to risk adjustment are described below.

Selecting Factors for Risk Adjustment

Patient factors selected for risk adjustment of a PRO-PM should be based on evidence that the factor affects the outcome independent of the treatment. Evidence would include peer-reviewed research literature, as well as clinical expert opinion. Informed patients can also provide valuable insights into potential factors that would affect an outcome. The factors, also known as covariates, may be different for different PRO concepts. For example, covariates associated with higher risk of pain might be severity of arthritis or time since surgery, while factors associated with a functional status outcome might include primary diagnosis, age, baseline functional status, and co-existing conditions.

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 present on admission, e.g., primary diagnosis, severity of illness, co-existing conditions, and baseline scores that affect outcomes, are also often included in risk adjustment models.

Baseline health status scores may be used in a regression model to adjust for the initial severity of the patients' status, such as functional status. Adjusting for baseline scores in regression models has been an area of discussion in the literature because the strength of the association between the baseline and follow-up (threshold) scores can affect analysis results.[50] For example, if a treatment is effective, and the outcomes of two groups are compared, the statistical significance of the follow-up score versus a change score (follow-up score minus the baseline score) will depend on the correlation between the baseline and follow-up score. 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 score. If the correlation between baseline and follow-up scores is low, then the follow-up score is more likely to be significantly different compared to the change score.[9, 50]

Psychosocial and other factors, e.g., motivation, understanding, engagement, adherence, and readiness to change, have been suggested as potential factors to include as covariates in regression models for PRO-PMs. The inclusion of psychological factors in performance measure risk adjustment models is controversial, and they are typically not included as a risk factor because an effective clinician can sometimes change a patient's status. For example, physical therapists at a certain facility 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. In addition, psychosocial data has not typically been available for patients, so use of these factors as covariates would likely require additional data collection.

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 both poorer outcomes and with disparities in care. These factors are not typically included in risk models for performance measures. Including factors associated with such disparities in risk adjustment models could mask quality problems due to disparities[43] 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."[3,43]

Many measures have not adjusted for these factors or stratified data by these factors. Providers with a higher proportion of disadvantaged patients may receive a lower performance score. 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 two 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.

Research examining the association between race/ethnicity and satisfaction with healthcare services has suggested that an expectation of outcomes plays a role in observed satisfaction differences.[51,52] These outcome expectation differences might best be addressed by clinicians and should not be risk adjusted.

Addressing Alternative Data Collection Sources, Modes, and Methods

If the source, mode, or method of data collection varies across persons, providers, or time, the provider-level performance measure data may not be comparable. Risk-adjustment methods may be appropriate to adjust for systematic differences tied to varying data-collection methods. If research comparing the alternative source, mode, or method shows that the PROM data are equivalent, then data could be pooled without adjustment regardless of data collection source, mode, or method. If, however, data are not equivalent, it is important to know if research has documented systematic differences across alternative sources, modes, or methods. If a systematic difference has been identified, adjustment factors could be part of the risk-adjustment model.[53,54] If research has found the scores are not equivalent or systematically different, data should not be pooled, and the data for selected patients would be missing for the performance measure. For the PHQ-9 instrument, a clinician observation version of the PROM has been developed, and initial validation testing has been conducted comparing it to the Cornell scale; however, equivalency of the self-report and clinician observation measures has not been established.[42,55] For the performance measures "Depression Remission within 6 Months," (NQF # 0711) and "Percent of Residents with Moderate to Severe Pain (Short Stay)" (NQF # 0676), 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,[56-60] 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, data 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 practices for translation and cultural adaptation of PROM instruments.[61,62] It is worth noting that there are now more than 75 translations of the PHQ-9 available (http://www.phqscreeners.com/overview.aspx).

Risk-Adjustment Methodology

There are several approaches to account for case-mix differences. 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 two or more groups. A second risk-adjustment approach uses regression modeling with demographic, clinical, 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, and tends to be somewhat homogeneous, regression modeling may be adequate to adjust for several covariates such as 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. As previously noted in the reliability section, hierarchical modeling with Bayes shrinkage estimators provides a strategy for improved PRO-PM estimates for providers that previously may have been too small and unstable to yield reportable measures. 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.[32]

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-PM that is not risk adjusted is the measure "Percent of Residents who Self-report Moderate to Severe Pain (Short Stay)" (NQF # 0676). 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.

Some providers may have a 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.

There is also a threat to the validity of the performance score if the identification of the target population is not consistent across providers or recruitment of the target population is inconsistent across providers.

Use of PRO-PMs in Different Programs

As noted above, measures of depression, pain, and functional status have relied on the patient's voice for many years to measure outcomes or change in status between the beginning of treatment and some subsequent point in time. In the United States, the advent of performance-based payment systems has led to the development of PRO-PMs that can be used to evaluate provider performance. One of the most advanced groups is the Minnesota Community Measurement Program, which is working with the state and other insurers to develop valid and reliable PRO-PMs that can be used for accountability purposes.

Since 2009, Minnesota has been requiring public reporting of a depression measure based on several NQF measures that rely on the PHQ-9: Depression Remission at Twelve Months (NQF #0710), Depression Remission at Six Months (NQF # 0711), and Utilization of the PHQ-9 (NQF # 0712)). These measures are being used in at least two pay-for-performance programs as an incentive in a physician payment program. Additional work is currently being done to develop threshold measures (percentage of patients whose responses indicate an improvement). However, many issues remain to be addressed that take into account remission and other factors that may mask performance when aggregated in this manner.

Additional measures used by Minnesota include an asthma composite measure. This measure has three components targeting asthma control, risk of exacerbation, and a process element documenting whether an asthma action plan is in place. The first two components have been in use for 2 years, and are reported on the public reporting website. The use of the last component in the composite measure is still being developed.

Minnesota also required through legislation that two types of orthopedic measures be collected: change in knee function following a total knee replacement and change in disability following lumbar spine surgery. The baseline data for knee replacement patients is collected before the surgery and then two times after the surgery (3 months and 1 year after surgery). The lumbar spine patient is assessed before surgery and 1 year after surgery. Selection of post-surgical assessment times was designed to coincide with existing followup appointments; however, this issue was extensively discussed, and it is not at this time a perfect match with the expected follow-up visits.

Minnesota is also working on a pediatric preventative depression measure. This will be a process measure, although the exact instrument and many of the methodological issues remain to be resolved. Differences in reliability between patient, parent, and proxy responses are particularly problematic with younger pediatric populations.

Other states are developing similar PRO-PMs but, as noted by Minnesota's progress, many issues must be considered before using a PROM within a performance measure. Many of the challenges are discussed in this paper, but they increase exponentially with measure implementation. The Quality Alliance Steering Committee, co-directed by the Agency for Health Research and Quality and the Brookings Institution, are working with performance-measure user communities to minimize some of the implementation issues that occur when moving from the research behind NQF's endorsement to the implementation at the organization or community level.

International Experiences

While many European countries are involved in PROM development, only a few are using the PROMs for performance measure and accountability. This section discusses some of the uses of PROMs outside the United States, including England and Sweden. Both countries have done extensive work in PROM development, but their use of the data is very different.

England

In 2005, England's National Health Service Department of Health commissioned a feasibility study of PROM data collection, which was conducted by the London School of Hygiene and Tropical Medicine. The study tested the feasibility of routine pre- and post-operative PROM data collection from patients undergoing elective surgical procedures and examined options for analyzing and presenting these data. The study participants were 2,400 patients treated at 24 centers who were scheduled for any of five elective procedures: unilateral hip replacement, unilateral knee replacement, treatment of varicose veins, removal of cataracts, and groin hernia surgery. Patients who were unable to complete the written questionnaire in English, due to cognitive impairment, poor sight, literacy or language issues, were excluded. All patients were asked to complete the first part of the EQ-5D and four groups were asked to complete a condition-specific measure: the Oxford Hip Score for hip replacement, the Oxford Knee Score for knee replacement, the Aberdeen Varicose Vein Questionnaire for varicose vein surgery, and the VF-14 for cataract removal. Patients undergoing groin hernia surgery were asked to complete the SF-36 (UK version 2).

The pilot offered several valuable lessons for England's National Health Service:

- Researchers recommended that patients be asked to complete the pre-operative surveys at the time of admission rather than during the pre-operative assessment.
- Staff training was recommended as a way to increase the percentage of patients asked to complete the pre-operative survey.
- Nursing staff tended to be better than clerical staff with administering the questionnaires.
- Overall, few patients in the pilot study were ineligible to complete the survey due to cognitive or language issues. However, among patients with cataracts, up to 30% were unable to participate due to vision problems.
- Thirteen percent of patients declined to participate.
- For patients with cataracts, the VF-14 showed only moderate responsiveness and concerns about the content validity of the PROM were expressed. The EQ-5D was unresponsive for patients undergoing cataracts surgery.
- The SF-6D was longer and had a higher incidence of missing data, so the researchers recommended use of the EQ-5D.
- An independent third party was recommended for post-operative data collection.

Since April 2009, providers offering four elective interventions have been required to collect and report PROM data as specified in the Standard NHS Contract for Acute Services. This means that all providers of NHS-funded unilateral hip and knee replacements and groin hernia and varicose vein surgeries are expected to ask patients undergoing one of these procedures to complete a pre-operative PROMs questionnaire. PROM data for patients undergoing cataract surgery is not collected, perhaps based on the feasibility study that showed challenges with pre-surgical data collection and concerns about the validity of the selected PROMs. Patients are asked to fill out the pre-survey after they have been medically cleared for surgery. Post-operative questionnaires are then sent to patients following their operation. The post-operative survey is sent out to patients 3 months post-surgery for hernia and varicose vein surgery and 6 months after hip or knee replacement surgery. The baseline and follow-up surveys are then linked with hospital episode data. Between April 2010 and March 2011, there were 245,516 eligible hospital episodes, with 171,499 (69.9%) pre-operative questionnaires returned. In response to the 171,499 pre-operative questionnaires returned, 162,614 (81.0%) post-operative questionnaires were sent out, and 131,696 (81.0%) were returned.

The Health and Social Care Information Centre (HSCIC) publishes data in the form of funnel plots on their website on a monthly basis. Provider-level comparative data identifies providers that are outliers, both positive and negative. The NHS is starting to link performance based on quality metrics to payment based on the Commissioning for Quality and Innovation framework (http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_091443). The providers work with the local health authority to select and define the performance measures; only one measure is currently based on a PROM. This program offers the potential for a payment increase of 1.5%.

England's experience with these PROMs and the public reporting of these data may be helpful to NQF's efforts. A review of data-collection procedures found that the biggest reason for missing data was the failure of staff to invite the patient to participate in the survey. For patients who cannot complete the survey, an interviewer is assigned to the patient. An independent third party collects the post-surgery data to improve response rates.

When aggregating the PROM data to the provider level, analysts impute missing data from survey items but not missing surveys. England has explored several options to deal with low response rates including the Heckman method, a statistical approach to addressing non-randomly selected samples,[63] but discussions are ongoing.

England has a goal of 80% recruitment rates and a minimum sample size of 150 cases. The exclusion criteria for PROMs in England included those patients who are 16 years of age or younger, are unable to complete the survey in English, or died prior to the scheduled post-surgery questionnaire. The NHS also prepares quarterly reports on mortality at the provider level. Patients must have completed a pre- and post-surgery questionnaire to be included in the performance measure.

The PROM data are risk-adjusted using the Generalized Least Squares fixed effects model. Risk adjustment variables include age, sex, ethnicity, pre-operative PROM scores, whether the patient had assistance in competing the survey, prior surgery, co-morbidities, and duration of the problem. Due to the high rate of missing data additional covariates such as body mass index, and surgeon-reported clinical severity measure were not used. The risk adjustment variables used vary by procedure. The case-mix adjustment process includes three steps:

- Estimation of the impact of the risk adjustment variables
- Generation of the patient-level predicted scores
- Aggregation to the provider level and case mix adjustment

When comparing data, the Department of Health in England uses statistical significance testing because the researchers felt that using the minimal clinically important differences has limitations. Statistical comparisons are based on 95 percent power and a significance level of $p < 0.002$. Comparisons with the national data are done using the funnel plot in the score-comparison spread sheet. If an organization's score isn't contained within the funnel, the organization is considered to be significantly different from the national providers.

Sweden

Sweden operates a decentralized national health care system. County councils manage the health care for their citizens at the local level, through elected officials. The national government of Sweden developed a framework for quality in 2009, and outcome data is collected and stored in more than 100 registries with approximately 75 of these registries including data from at least 1 PROM. The registry data is used to improve clinical care and for research. The Swedish National Board of Health and Welfare and the Swedish Association of Local Authorities and Regions require the use of PROMs and publish select performance measures. Sweden distinguishes PROMs from patient-reported experience measures (PREMs), which focus on patients' perspective on their care. The report *Quality and Efficiency in Swedish Health Care –Regional Comparison*, published every year for the past 5 years, focuses mostly on regional comparison of medical outcomes, with limited provider-level data. The focus has been on patient outcomes, availability of services, and cost. The report was designed for both transparency and management of the healthcare system.

The recent *Quality and Efficiency in Swedish Health Care –Regional Comparison 2010* contains 134 indicators, of which 8 are PROMs:

- Patient-reported complications after hysterectomy
- Patient-reported complications after uterine prolapse surgery
- Patient-reported outcomes of total hip arthroplasty

- Patient-reported improvement after initiation of biological drug therapy for rheumatoid arthritis
- Patient-reported improvement after initial care for rheumatoid arthritis
- Activities of daily living ability 3 months after stroke
- Satisfaction with stroke care at hospitals
- Patient-reported outcome of septoplasty

The report includes response rates, and data are stratified by male and female. PROMs that have existed over a longer period appear more robust in their reporting. An example is the PROM "Satisfaction with Stroke Care at Hospital." The National Stroke Registry mails a questionnaire to the patient and family 3 months after the patient's stroke. While the overall response rate for this measure was close to 90%, regional response rates varied from 80% 100%. At the provider level, response rate varied from 53% to 100%. The PROM "Activities of Daily Living Three Months after Stoke" was one of the few PROMs that was risk-adjusted, and covariates for this measure are age and level of consciousness on arrival at the hospital.

The 2010 report provides a list of attributes used to select measures, though not specifically PROMs:

- Quantifiable and available at a national level.
- Generally accepted and valid, including face validity
- Relevant, including volume and cost
- Amenable to interpretation
- Capable of being influenced
- Outcome or process measure.

"Patient-reported Outcome Measures and Health-economic Aspects of Total Hip Arthroplasty: A study of the Swedish Hip Arthroplasty Register"[64] does a more extensive review of the Registry experience and the data it has been collecting since 2002 on PROMs. The report emphasizes the need for a PROM to have the following characteristics:

- It should be valid and reliable.
- It should combine generic and disease-specific data to gain a better perspective of the patient's overall health and the impact of a specific condition. A generic measure of health would also allow for comparison of different patients as well as populations.
- The number of questions in the survey should be limited to ensure a high response rate.

- It needs to detect change.

Sweden generally has a high response rate to PROMs, and this is attributed in part to unique patient identifiers that allow patients to be tracked. The Registry tested both an internet version and pen-and-paper collection system; the web-based application did not improve response rates. Timing of the post-surgery survey may also be important. Comments in the Swedish Hip Arthroplasty Registery report suggest that England's approach of a post–surgery survey 6 months after a hip replacement may be too soon after the surgery.

Sweden has a long history of collecting PROMs data, and their use of the data for research and public health activities has evolved:

- The use of PROMs has accelerated with a national requirement to included PROMs in national registries.
- The PROMs include both disease-specific instruments and the generic measure EQ-5D.
- The response rate is high at a national level. The Registry works with individual providers to improve their response rates. Sweden does attribute the high response rate to limiting the length of the PROMs.
- PROMs can help identify populations that did not have the expected outcomes, warranting further research on these sub-populations.

eMeasures

Electronic measures, or eMeasures, are electronic versions of performance measures. They are based on EHR and require standardized format specifications. Some work has been done by NQF and others in the health information technology field to build on the directives of the Health Information Technology for Economic and Clinical Health (HITECH) legislation to create data standards that can be used to transfer data across electronic platforms. However, many of these standards are still being developed. SNOMED, LOINC, and a few other classification systems exist, which the HITECH community and others are building on. However, eMeasures first require consensus on the content of the performance measure, and then standard electronic terminology must be developed to submit the measure electronically or exchange the data.

Building eMeasures from the electronic PROMs requires several additional steps. First, exact specifications of the performance measure must be developed. This requires use of common data standards to specify both the data source and the potential responses. Second, these standards must be incorporated into the EHR to allow electronic transfer of the data elements into a performance measure. Incorporating standard performance measures in these data systems will become easier as data standards are developed for more concepts.

In general, one of the biggest challenges in moving toward eMeasures is maintaining the intent of the measure throughout the process. For ePRO-PMs, this is complicated by several factors associated with patient reporting approaches. In addition to the reliability and validity concerns noted above in developing a PRO-PM, advances in electronic communications open the door to numerous reporting approaches for PRO-PMs, and each may introduce error into the provider-level measure. First, patient-reported data may be submitted in several ways, including in-person visits, telephone-based responses, and e-mail responses. Using the EHR as the source for the eMeasure would require that the patient-reported information be captured in the EHR so the data can be transferred into the performance measure calculation. As noted above, each of these reporting approaches may affect the reliability and comparability of individual responses across providers. These issues are not different from those raised above in highlighting the impact of administration modes on data reliability, but they are multiplicative in their potential impact if standard protocols are not used to capture the patient responses. This issue can be minimized by giving the patient a survey to complete (on paper or electronically). However, the proliferation of Smart phones, remote monitoring devices, application development platforms (e.g., iPhone and iPad apps), and other networks are enabling extensive opportunities for variations in how patients can report data.[65] These opportunities may reduce the standardization of the responses at the initial stage.

Second, some types of patient-reported data may need to be treated as confidential (e.g., satisfaction with provider, adequacy of the medications instructions). These types of measures are typically aggregated and shared at the plan or provider level. Their use in electronic measure specification would require they not be incorporated in the patient record in order to avoid privacy concerns and maintain the reliability of the collected data.

Clinicians and key stakeholders from the NQF's eMeasure Learning Collaborative have identified three key success factors in developing eMeasures that can be applied, with special consideration, as broad guidelines to PROs as well. These factors are business, function, and content.

Business relates to the approach to standardize the data submission. It will be important to use standard protocols for collecting the information electronically. This may require supervision or facilitation of the activity by a clinician or staff member to ensure standard approaches. Electronic systems and versions must be consistent across settings for data transfer to be fluid.

Function should be considered in terms of the ultimate use of the data for quality measurement. One way measure developers have addressed this is to develop measures around critical path areas or issues considered important to the patient or the clinician in restoring health. These critical paths can be defined by the patients in developing future performance measures. For example, one of the critical path measures in cardiovascular care is the patient-safety measure for inpatient infusion pump therapy. This measure is intended to address the 30–65 percent of medical errors involving a pump of some sort that are related to an adverse event. But additional critical paths can be defined by the patients' perspective to develop performance measures that are important to them. This will help strengthen the eMeasure interpretation. Many of NQF's identified high-impact conditions could benefit from the subjective nature of electronic patient-reported outcomes, such as major depression. Patients may be more comfortable and subsequently more truthful reporting information without a clinician present and entering data into a computerized system that they feel is secure.

One caution in this area is the need to focus on patient privacy. Although patient-reported outcomes are useful to include in their EHR, other types of patient-generated performance measures may be less appropriate to store in the record. Many of the measures may be based on patient perception of care or satisfaction. Collection of this data is often based on anonymity or aggregation at the provider level without providers being able to identify individual respondents. This privacy will need to be protected as these performance measures are developed.

The third area raised by the NQF advisory group is content. Content is the consideration with the most implications in developing an eMeasure for PROs. To start, it deals with key factors in PROs: information capture and subjectivity. When collecting data, especially patient-reported data, it is critical that appropriate vocabulary is both available and is properly utilized and interpreted by the patient and clinician. A standardized, clear, and consistent language is a challenge to establish across settings with heterogeneous patients, but it is crucial in compiling a useful data set. In addition to this, an item must have a delicate balance between accommodating patient comprehension levels and capturing data that is detailed enough to take advantage of the output capabilities of electronic systems. Measure developers should consider the larger context of measure families and the operationalization of a measure as part of a composite group, as well as the removal of a measure from certain inappropriate settings. For example, a patient's goal of returning home may be part of a composite for a skilled nursing facility, but this may not be an appropriate item for a resident living in a long-term care facility. The data elements must be able to be inserted electronically into and removed from use in various settings with ease.

It is pertinent to note that eMeasure success is more dependent on the people who report than on the actual technology. An excellent electronic system is useless with a low utilization rate. The measures should be easily integrated into the patients' and clinician's routine, so as to reduce burden. One final challenge to consider when developing a PRO eMeasure is to capture red flags sufficiently. While it is important to automate a measure system to identify outliers, we must consider how far it is reasonable to go in predefining possible scenarios with the system. Overall, key considerations for eMeasure development should all aim to ensure that original clinical intent is captured and reported for the purposes of its use.

Conclusion

The area of PRO-PMs is relatively young and still evolving. Although the patient's voice is often included in experience with care measures, it is incorporated into the science of PRO-PMs much less frequently. Some work on pain management and depression monitors their effects on quality of life, but much more remains to be done in using other PROMs as performance measures. The importance of the underlying science is critical as one moves from measuring outcomes at the individual level to holding organizations accountable for patient outcomes. As noted in the discussions above, many issues need to be addressed when developing performance measures, and these can be complicated when using PROMs for accountability purposes.

This paper discusses the key issues that must be considered when developing and evaluating PRO-PMs. Many factors may affect the scientific integrity of PRO-PMs. Some critical questions and key considerations for moving forward in using PROMs for performance-measure development are:

- Is the PROM-PM's instrument reliable and valid for the target population?
  - Considerations: It is a necessary but not sufficient requirement for the PROM to be reliable and valid before it can be used in a performance measure.
- Is the PRO-PM 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 random error at the individual level?
  - – Is the sample size adequate for providing robust results?
  - – Have any adjustments been made to reduce random error that may lead to misleading results?
  - Considerations: Until there is substantial adoption of PROMs, initial reliability testing will be limited. Reliability studies should be planned to include a number

of providers each with sufficient numbers of patients so that variation between and within providers can be analyzed.

- Is the PRO-PM construct valid? Does it permit inferences about the quality of care in the organization?
  - Are the results clinically meaningful?
  - Are these PRO-PMs important to the patient?
  - How has validity been tested?
    - Does it meet face validity requirements?
    - Are the desired patient outcomes consistent with clinical expectations for that type of case?
  - Considerations: The performance measure must have face validity to be scientifically acceptable. Although additional validity testing is desirable, without substantial use of PROMs in multiple providers, initial validity testing will be limited or may not be possible. To meet standards of robustness, validity studies need careful planning to have sample sizes large enough to test. This may not be possible until a measure is in wide use, which typically comes after, rather than before, endorsement.
- How is the PRO-PM performance score calculated?
  - When is a change score more (or less) appropriate than a threshold value for an expected outcome?
  - Are patient preferences taken into account and, if so, how?
  - Considerations: Patient preferences may be very important when deciding whether to use a change score or threshold value for an expected outcome. For outcomes that rely on patient perceptions and preferences, such as the definition of acceptable pain levels (relative to increased medication or other factors), it may be more effective to use threshold values. Certain types of cases may be limited in such a way that their expected changes from the treatment are limited, and these types of performance measures may rely on threshold measures rather than change as a performance metric. The decision should be consistent with the existing scientific evidence.
- What risk adjustment techniques are being considered or used?
  - Are the covariates appropriate for the outcome being measured? Do the factors affect the outcomes, independent of the treatment provided?
  - Does the method control for test effects associated with sources of data, the methods used, or the mode of administration, which may affect provider PROM scores?

    - Considerations: Risk adjustment is an important step to ensure fair comparisons across providers. However, caution should be used to ensure that the methodology does not adjust for factors for which providers should be held accountable.

- What other factors threaten the validity of the performance measure?
    - Missing data?
    - Inadequate risk adjustment?
    - Considerations: Sample robustness and other considerations will be important as the patient voice becomes incorporated into performance measures. Because patient responses are voluntary, these types of measures may have greater problems of robustness than clinically assessed outcomes measures.

Few answers to the questions above are right or wrong; the best approach will likely depend on the PROM and the goal of the performance monitoring. For example, using PRO-PMs to pursue internal quality-improvement goals may be considered very important. By contrast, using PRO-PMs to meet external regulatory requirements of ensuring at least a minimal level of quality may weigh these factors differently and not give as much attention to patient preferences until they can be proven to be robust measures.

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.[66] This may require 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-PMs 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 CMS will help lay the groundwork for incorporating these items into daily treatment, workflow processes, and electronic records.

PROMs rely on the patients' assessment of their health status, a subjective measure that may be affected by the patient's expectations and other perceptions that may differ from the clinicians' expectations. This distinction may vary in its impact on performance measures, depending on whether the PROM was asking the patients to report on actual performance or on their evaluation of the experience. The patient voice is important, particularly as it provides the best measure of whether the persons seeking care regarded their treatment to be effective. In some cases, it also may reflect less realistic or achievable goals than the clinicians' perspectives. Both the patients and the clinicians' perspectives are important in measuring performance.

Moving the patient’s voice into clinical practice is key for the future of health care delivery that is person-centered. As noted by the Institute of Medicine in *Crossing the Quality Chasm*,[67] 6 of the 10 recommendations for improving the quality of the healthcare system address direct involvement of patients in their own care. Engaging the patients in the process of care, particularly by noting their outcomes, is key to developing better outcomes and therefore improving health. 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. Cella D, Hahn EA, Jensen SE, Butt Z, Nowinski CJ, Rothrock NE. . *Methodological issues in the selection, administration and use of patient-reported outcomes in performance measurement in health care settings.*2012.

2. National Quality Forum. *National Voluntary Consensus Standards for Patient Outcomes.* Washington, DC.2009.

3. National Quality Forum. *Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties.* Washington, DC.2011.

4. Kane RL, Radosevich, D. M. *Conducting Health Outcomes Research*. Sudbury, MA: Jones & Bartlett Learning; 2011.

5. Iezonni L. *Risk Adjustment for Healthcare Outcomes.*2003.

6. McGlynn EA, Asch SM. Developing a clinical performance measure. *American journal of preventive medicine.* Apr 1998;14(3 Suppl):14-21.

7. Donabedian A. Evaluating the quality of medical care. 1966. *The Milbank quarterly.* 2005;83(4):691-729.

8. 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.

9. Streiner DL, Norman GR. *Health measurement scales : a practical guide to their development and use*. 4th ed. Oxford: Oxford University Press; 2008.

10. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. *Mayo Clinic proceedings. Mayo Clinic.* Apr 2002;77(4):371-383.

11. Collins LM, Johnston MV. Analysis of stage-sequential change in rehabilitation research. *American Journal of Physical Medicine & Rehabilitation.* 1995;74(2):163-170.

12. Stineman MG, Henry-Sanchez JT, Kurichi JE, et al. 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(2):126-140.

13. Stineman MG, Maislin G, Fiedler RC, Granger CV. A prediction model for functional recovery in stroke. *Stroke; a journal of cerebral circulation.* Mar 1997;28(3):550-556.

14. Stineman MG, Ross RN, Fiedler R, Granger CV, Maislin G. Staging functional independence validity and applications. *Archives of Physical Medicine & Rehabilitation.* 2003;84(1):38-45.

15. 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(1):29-37.

16. 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(3):454-462.

17. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank quarterly.* 2004;82(4):661-687.

18. 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(3):E191-196.

19. Adams JL. *The Reliability of Provider Profiling: A Tutorial*: RAND Corporation;2009.

20. Adams JL, Mehrotra, Ateev, McGlynn, Elizabeth A. *Estimating Reliability and Misclassification in Physician Profiling*: RAND Corporation;2010.

21. Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: the importance of reliability adjustment. *Health services research.* 2010;45(6 Pt 1):1614-1629.

22. 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(2):231-235.

23. Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. *International Journal for Quality in Health Care.* Dec 2001;13(6):481-488.

24. Ash A, Fienberg SE, Louis TA, Normand S-LT, Stukel TA, Utts J. *Statistical Issues in Assessing Hospital Performance* January 27, 2012.

25. 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.* Apr 2009;47(4):378-387.

26. 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.* Jun 9 1999;281(22):2098-2105.

27. Roebroeck M, Harlaar J, Lankhorst G. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. *Physical therapy.* 1993;73(6):386-395.

28. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. *Physical therapy.* Jun 1993;73(6):386-395; discussion 396-401.

29. Austin PC. The reliability and validity of Bayesian measures for hospital profiling: a Monte Carlo assessment. *J Statist Plann Inference.* 2005;128(1):109-122.

30. National Quality Forum. *Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety—Composite Measures: A Consensus Report. .* Washington, DC.2009.

31. Normand SL, Glickman ME, CA G. Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. *Journal of the American Statistical Association.*92(439):803-814.

32. Ash AS FS, Louis TA, Normand S-LT, Stukel TA, Utts J. *Statistical issues in assessing hospital performance*: CMS;2012.

33. Fitch K. *The Rand/UCLA appropriateness method user's manual*. Santa Monica: Rand; 2001.

34. 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.* Apr 5 2005;45(7):1147-1156.

35. Hibbard J, Pawlson LG. Why not give consumers a framework for understanding quality? *Joint Commission journal on quality and safety.* Jun 2004;30(6):347-351.

36. Basch E. The missing voice of patients in drug-safety reporting. *The New England journal of medicine.* Mar 11 2010;362(10):865-869.

37. Tulsky DS, Kisala PA, Victorson D, et al. Developing a contemporary patient-reported outcomes measure for spinal cord injury. *Archives of physical medicine and rehabilitation.* Oct 2011;92(10 Suppl):S44-51.

38. 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.* Oct 2009;18(5):374-379.

39. Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine.* 2006;21(6):547-552.

40. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry.* 2008;8:46.

41. Yang FM, Jones RN. Center for Epidemiologic Studies-Depression Scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *Journal of Clinical Epidemiology.* 2007;60(11):1195-1200.

42. Saliba D, Buchanan, Joan. *Development and validation of a revised nursing home assessment tool: MDS 3.0.*: RAND Corporation;2008.

43. National Quality Forum. CSAC Guidance on Quality Performance Measure Construction: National Quality Forum,; 2011.

44. Johnson TP, Wislar JS. Response rates and nonresponse errors in surveys. *JAMA : the journal of the American Medical Association.* 2012;307(17):1805-1806.

45. 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(12):2381-2386.

46. Klein DJ, Elliott MN, Haviland AM, et al. Understanding nonresponse to the 2007 Medicare CAHPS survey. *Gerontologist.* 2011;51(6):843-855.

47. Dillman DA, Smythe HD, Christian LM. *Internet, Mail and Mixed_mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley and Sons, Inc; 2009.

48. Hays RD, Kim S, Spritzer KL, et al. Effects of mode and order of administration on generic health-related quality of life scores. *Value in Health.* Sep 2009;12(6):1035-1039.

49. 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(3):833-848.

50. Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ.* Nov 10 2001;323(7321):1123-1124.

51. Cho H, Kim WJ. Racial differences in satisfaction with mental health services among victims of intimate partner violence. *Community Mental Health Journal.* 2012;48(1):84-90.

52. Hepinstall MS, Rutledge JR, Bornstein LJ, Mazumdar M, Westrich GH. Factors that impact expectations before total knee arthroplasty. *Journal of Arthroplasty.* 2011;26(6):870-876.

53. Skolarus LE, Sánchez BN, Morgenstern LB, et al. 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(3):510-515.

54. Coons SJ, Gwaltney CJ, Hays RD, et al. 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.* Jun 2009;12(4):419-429.

55. 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.

56. 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.* Dec 2010;69(6):525-532.

57. Chang PC, Yeh CH. Agreement between child self-report and parent proxy-report to evaluate quality of life in children with cancer. *Psycho-oncology.* Feb 2005;14(2):125-134.

58. Matza LS, Secnik K, Rentz AM, et al. 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.* Apr 2005;14(3):735-747.

59. 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.

60. Varni JW, Stucky BD, Thissen D, et al. 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(11):1109-1119.

61. 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(3):151-156.

62. Wild D, Grove A, Martin M, et al. 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.* Mar-Apr 2005;8(2):94-104.

63. Heckman J. Sample selection bias as a specification error. *Econometrica.* 1979;47(1):153-161.

64. Swedish Hip Arthroplasty Register. *Patient-reported Outcome Measures and Health-economic Aspects of Total Hip Arthroplasty: A study of the Swedish Hip Arthroplasty Register* 2010.

65. Shapiro M, Johnston D, Wald J, Mon D. *Patient-Generated Health Data.* RTP, NC: RTI International;2012.

66. Wu AW, Snyder C. Getting ready for patient-reported outcomes measures (PROMs) in clinical practice. *HealthcarePapers.* 2011;11(4):48-53; discussion 55-48.

67. Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Centurey*. Washington, DC: Institute of Medicine; 2001.

Appendix A

The American Association for Public Opinion Research (AAPOR) Council uses the following formulas for calculating response rates, including 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
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.

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