NATIONAL QUALITY FORUM

Measure Sets and Measurement Systems: Multistakeholder Guidance for Design and Evaluation

FINAL REPORT
JULY 31, 2020

qualityforum.org

Contents

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Foreword

We are at a critical juncture in the transformation of how we pay for healthcare. The increasing use of measures in value-based models of care is a core component of the next era of performance measurement. The COVID-19 pandemic has clearly demonstrated again the limits of fee-for-service payment and the critical need for population health-driven payment models. The required advancement in measurement inspires the need for multistakeholder, consensus-driven review—not only of individual measures, but also of how they are used together to judge quality performance in specific settings, by certain conditions, and by episodes of care.

To this end, NQF convened a Technical Expert Panel with 25 diverse experts from across the healthcare community to define measure sets and measurement systems, and establish key elements for their review and evaluation. We brought together patients and patient advocates, purchasers, public and private payers, clinicians, provider groups, measure developers and implementers, statisticians, and health services researchers with the common goal of ensuring that measure sets and measurement systems produce results that are accurate and actionable. Since we started this work over a year ago, we have released several related publications to ensure all of our stakeholders are involved and can keep up with our progress. It is also important to note that this work is funded by our members—a sign of its critical strategic importance to NQF and the ecosystem broadly.

The identification of common elements that define measure sets and measurement systems lays the foundation for future multistakeholder review. Building on our recent **Hospital Star Rating Summit**, recommendations from the **NQF-convened National Quality Task Force,** and our tradition of measure endorsement and selection, measure set and measurement system assessments can help to ensure that scientifically sound information is provided to patients and consumers to make informed healthcare decisions. Accurate assessments also help to ensure entities being measured have clear improvement opportunities and that higher observed performance is based on the delivery of higher quality care. Taken together, this work can help to drive our healthcare system to higher levels of performance.

The effort to establish standards for measure sets and measurement systems potentially represents the next step in the evolution of NQF's work to set standards for quality measures based on evidence and innovation to make care better for all people. This will particularly be essential as the ecosystem contends with the lasting impacts of COVID-19. We are indebted to our members, partners, and Technical Expert Panel for supporting this work, and for demonstrating yet again that NQF is the forum for tackling emerging national health needs that demand comprehensive approaches across sectors. Following the release of this final report, we expect to continue the next phase of work through strategic partnerships and pilot testing of these ideas. We also expect to look at broader strategic issues and questions underlying endorsement to ensure this program meets today's healthcare needs. We look forward to continuing our collaboration on the future of quality measurement.

Sincerely,

Shantanu Agrawal, MD, MPhil, President and CEO

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Executive Summary

Today's healthcare measurement landscape uses many individual measures to assess the quality of care and identify opportunities to drive improvement. Individual measures, however, are increasingly being used together to make broader inferences about quality and inform consumer decision making. While guidance to determine the rigor of individual measures is well-defined, as of now, there is not an established approach to assess the methodology of how measures used together determine performance results.

Based on input from a multistakeholder Technical Expert Panel (TEP), NQF proposes definitions and elements of measure sets and measurement systems, and puts forth an approach to determine if they are of sound design based on their intent. A "measure set" is defined as a group of individual measures that address an aspect of quality or cost, created for a specific purpose. A "measurement system" is a group of measures that, based on a predefined methodology, work together to assess quality or cost in relationship to a goal. These related-yet-distinct concepts are compared using examples, and the components of each concept are described in an effort to move the field toward a more unified understanding of their design. Elements have been synthesized in concrete sections to help readers visualize the progression of considerations that inform measure set and measurement system design. Topics and themes discussed in this report are connected and interdependent. As such, the concepts may not be fully represented by independent clauses or sections.

Well-defined elements of measure sets and measurement systems are essential to the establishment of a process to assess the soundness of their design. A consensus-based, standardized process that brings together multiple stakeholders to discuss measure sets and measurement systems fills a gap in the measurement landscape, which currently focuses on individual measures. The relationship between NQF's attempt to develop a method to assess measure sets, measurement systems, and ongoing NQF programs is acknowledged in this report.

With quality measurement becoming more complex, there is a need to ensure clarity and appropriateness in the way measures are used together to interpret quality. NQF welcomes partner organizations in advancing healthcare quality by ensuring performance measures comprehensively drive improvement in outcomes and reward high quality care.

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Introduction

After the release of the seminal works, *To Err is Human*[1] and *Crossing the Quality Chasm*[2], the quality enterprise began its improvement journey by defining individual measures that can be used to assess safety and quality. These measures have helped propel a movement that values a culture of safety and improves the quality of care for patients.

Measures are used in a number of ways to drive improvements in the healthcare system, as part of public reporting, pay-for-performance, and value-based programs. Publicly reported quality data also help inform consumers about where to seek medical care.[3,4] Public-facing quality programs like the CMS Overall Hospital Quality Star Ratings compile performance on quality measures across seven domains, consolidating these measures into a single summary score, from 1 star to 5 stars.[5] As such, there exist limitations in clearly defining what these scores mean. Payers and purchasers are also increasingly relying on measured quality performance to determine provider and hospital reimbursement, and to incentivize improvement in health outcomes. With the increasing cost and burden of measuring and reporting quality, and the need for measurement to reveal actionable opportunities for quality improvement, efforts need to thoughtfully advance the current state of healthcare measurement.[6] As the nation's healthcare delivery system transitions to value-driven models of care, quality measurement must support a comprehensively informed view of quality and more aggressively drive measure alignment across stakeholders.

Measurement systems themselves can face challenges such as limited data and measures, lack of robust audits, varied methodology in the creation of composite measures, comparing a diversity of provider and hospital types, and a lack of formal peer review of methods.[7] As measurement systems intend to assess and publish aggregate quality scores to help distill and ease communication of complex quality measurements, the design decision to support these goals should support the intent.

While healthcare quality measurement started to assess and improve health outcomes, today the measurement landscape faces both collective and evolving gaps as well as oversaturation in some measure areas. There are differing opinions of the value of different measure types (e.g., structure, process, or outcome) and debate between using more accessible administrative data (e.g., claims data) versus more detailed clinically enriched or patient-reported data. The current state of healthcare quality is marred by related measures that may not always work in sufficient synchrony to drive comprehensive and comparable quality performance results. Thus, the measurement enterprise still has opportunities to be more agile in driving innovation in healthcare quality.

MOVING BEYOND INDIVIDUAL MEASURES

While NQF evaluates the scientific merits of individual measures and provides guidance on their use, there is no established process for assessing how measures work together. Measures are often grouped into measure sets and then systematically used as part of a measurement system to evaluate quality in relationship to a goal. The way in which measures are aggregated together in a measurement system affects provider performance independent of a change in performance on an individual measure. Increasing use of measure sets and measurement systems for accountability and payment necessitates greater transparency and multistakeholder input. Yet, there is no consensus on the definitions or components of sets and systems and, often, there is a lack of clarity and consistency in the way measures are used together to make inferences about quality.

Measure sets and measurement systems should provide valid assessments of quality and reliable results to drive performance improvement, appropriately influence payment, and empower patients and other users to make more informed

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

healthcare decisions. NQF has increasingly led initiatives to further these goals. In phase 1 of our work, NQF created **an initial framework**[8] outlining definitions of measure sets and measurement systems. Building on this foundational work, NQF recently applied concepts from the framework to convene the **Hospital Quality Star Rating Summit**[9] and provide concrete recommendations to strengthen the reporting program. The Summit is an example of how a multistakeholder review of a measurement system can drive transparency and assess how performance measures are used together to support inferences about differences in provider quality performance.

In phase 2 of the measure sets and measurement systems project, NQF convened a TEP to review the initial framework, refine the elements of sets and systems, and help establish guidance for their design and evaluation. With the TEP's input, NQF has defined components of sets and systems that should be transparent and developed standardized multistakeholder approaches to assess their scientific appropriateness. NQF intends to test these approaches by engaging developers and multiple stakeholders in a consensus-driven process to help ensure measure sets and measurement systems are of sound design.

Project Design

In 2019, NQF convened a TEP of 25 members to discuss definitions, best practices, data issues, and unintended consequences of measure sets and measurement systems. TEP members included representation from a variety of stakeholders—patients and patient advocates, purchasers, public and private payers, clinicians, provider groups, measure developers and implementers, statisticians, and health services researchers. TEP members also had experience in performance-based payment, population health, and healthcare disparities. For a full list of TEP members, please refer to Appendix A: Technical Expert Panel Rosters and NQF Staff.

The TEP was divided into two subgroups: one focused on sets and another focused on systems. The full TEP participated in an initial orientation meeting in June 2019, where the group reviewed project objectives and preliminary definitions of sets and systems. In July and August 2019, the sets and systems subgroups met twice to discuss the elements of measure sets and measurement systems separately. The subgroups then reconvened as the full TEP and met monthly from September to November 2019 to discuss the alignment of elements and differences between sets and systems as well as review draft submission forms that NQF could use to evaluate sets and systems.

Defining and Designing Measure Sets and Measurement Systems

Initial efforts have demonstrated the importance of transparency and multistakeholder evaluation for how measures are used in sets and systems. Many of the elements of a measure set and measurement system are not clearly defined, nor are they evaluated for their scientific properties by a multistakeholder body, but they are used to assess provider performance. The way measures are used together may have a significant impact on how a provider is judged and how patients make choices about how and where to receive care. Our work demonstrates that several elements of sets and systems should be made transparent and evaluated to ensure appropriate scientific methods are used for development and that results accurately reflect quality of care. NQF puts the definition and elements of sets and systems out to the field to characterize their significance and draw attention to this important measurement and implementation opportunity.

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

RELATIONSHIP BETWEEN SETS, SYSTEMS, AND COMPOSITE MEASURES

There is an interrelationship between measure sets and measurement systems, evident from several common elements across both. While a pathway exists that links individual measures to sets and then to systems, for the purposes of design and evaluation, there are distinctions to consider:

1. Measure sets may be designed to fit more than one system. Ideally, a measure set would be thoroughly vetted and maintained so it could be used in different systems without an expectation of changing its construct.
2. A pre-established method to determine performance of entities relative to one another is not an inherent characteristic of a measure set—that remains a distinct aspect of a measurement system. Since measure sets may be applied to more than one system, similar to individual measures being used for multiple purposes, there may be value in distinct evaluation processes. For measure sets, assessment would determine if the set, collectively, is appropriately aligned with the purpose for which it was developed.
3. Generally, measurement systems contain a measure set plus other components (e.g., method for how the measures will be aggregated, incentive structure). Note that all sets and systems may not have all of the same components. For example, a system might not risk adjust the final quality assessment or assign measures to groups, but, if applicable, each of these elements should be considered and should be made transparent. When designing or evaluating a measurement system, its measure set would be considered within the context of the specific system. There is currently no agreed upon approach to evaluate how a measurement system's design aligns with its goal—a gap in the measurement infrastructure this work addresses.

While there is not yet widespread agreement in the field regarding the definitions of measure sets and measurement systems, composite performance measures are well-defined. A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.[10] A composite performance measure is a measure set plus an established method to combine measure performance into a single score. However, a measure set does not have to be a composite, meaning it does not have to roll up into one score using an inherent scoring method. Measure sets made up of measures from different developers often do not have an accompanying scoring method that would be universally used across multiple measurement systems.

A composite performance measure can be differentiated from a measurement system based on its context. A measurement system has a broader goal and an associated incentive structure. While many measurement systems involve the aggregation of a measure set to a single score, an individual measure or individual measures with benchmarks and a mechanism that incentivizes entities to achieve certain performance thresholds could also form a measurement system.

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

EXAMPLES OF MEASURE SETS

Measure Set	Rationale
Core Quality Measures Collaborative (CQMC) Core Sets	"The CQMC defines a core measure set as a parsimonious group of scientifically sounds measures that efficiently promote a patient-centered assessment of quality and should be prioritized for adoption in value-based purchasing and alternative payment models (APMs)."[11] The CQMC core sets are designed to be implemented in various measurement systems across both commercial and government payers.[12] There is no specific methodology for how the measures should be aggregated or how clinicians should be scored using these measures—a differentiating factor of sets versus systems.
Rural Health Core Set	The Rural Health Core Set is made up of the best available rural-relevant measures to address the needs of the associated population. Certain criteria were used for measure selection (e.g., NQF-endorsed, cross-cutting, resistant to low case-volume, addressing transitions in care, and addressing priority conditions for the rural population). Many of the measures in the core set generally may be suitable for use in various measurement systems (e.g., CMS reporting programs). The Rural Health Workgroup did not seek to select measures for any current or future program.[13]
ORYX® Measure Set	"ORYX® is a set of performance measures required by The Joint Commission. Hospitals seeking accreditation from The Joint Commission must submit some combination of ORYX® measures to fulfill the requirements. The measures are also meant to support organizations in their quality improvement efforts."[14] Acute Care Hospitals, Freestanding Psychiatric Hospitals, Critical Access Hospitals are each required to submit a different combination of electronic clinical quality measures (eCQMs) and chart-abstracted measures based on volume and average daily census.[15]

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

EXAMPLES OF MEASUREMENT SYSTEMS

Measurement System	Rationale
Overall Hospital Quality Star Ratings	The overall hospital quality star ratings summarize a variety of measures across seven areas of quality—Mortality, Safety of Care, Readmission, Patient Experience, Effectiveness of Care, Timeliness of Care, and Efficient Use of Medical Imaging – into a single star rating for each hospital.[16] This is a measurement system as it includes a measure set of 51 measures grouped into 7 different areas and includes methodology for aggregating the measures and scoring hospitals.
Merit-based Incentive Payment System (MIPS)	"MIPS was designed to tie payments to quality and cost efficient care, drive improvement in care processes and health outcomes, increase the use of healthcare information, and reduce the cost of care."[17] MIPS uses measure sets, but functions as a pick list in which providers can select any six measures to report or choose to report an entire specialty measure set. MIPS has a predefined methodology to determine a final score for reporting clinicians. Four performance categories make up a final score, which determines payment adjustment.[17]
Medicare Shared Savings Program	"The Shared Savings Program is committed to achieving better health for individuals, better population health, and lowering growth in expenditures."[18] In this voluntary system, an Accountable Care Organization (ACO) is held accountable for the quality, cost, and experience of care of an assigned Medicare fee-for-service (FFS) beneficiary population. A measure set plus a predefined methodology determine ACO performance. Measures used in the program focus on patient/caregiver experience, care coordination/patient safety, preventive health, and at-risk populations. To calculate an ACO's quality performance score, four measure domains are weighted equally at 25 percent. For pay-for-performance measures, performance against benchmarks corresponds to the amount of quality points an ACO earns for each measure. ACOs also earn points for complete reporting of pay-for-reporting measures and can earn additional points for each domain in which they show significant improvement.[19,20]

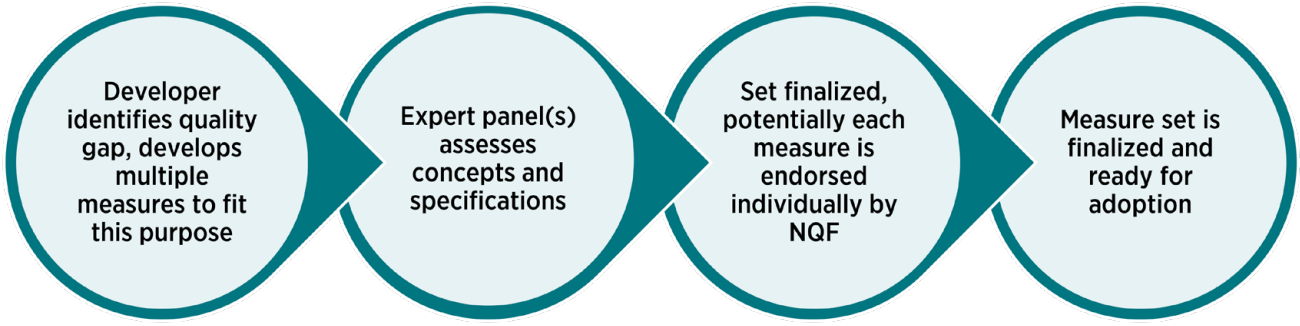
© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

MEASURE SET DESIGN ELEMENTS

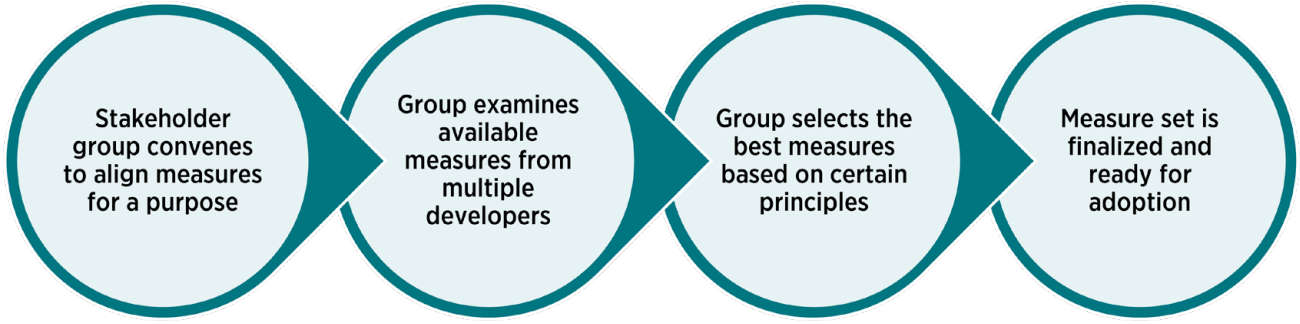
A measure set is defined as a group of individual measures, created for a specific purpose, that address an aspect of quality or cost. Developing measure sets is a strategy to comprehensively assess quality for a particular topic. Using measure sets may also reduce measurement burden by promoting implementation of the same measures—those deemed most valuable across users.

Shown below are two distinct pathways commonly used to create measure sets: internal development by a single developer or external development by an organization. The design guidance and measure set element descriptions are applicable to both cases. Both pathways can result in a measure set that would be used in one or more measurement systems. While these pathways appear linear, they are somewhat continuous as measure sets need to be maintained over time. Stakeholder feedback and patient engagement are also integral components of the creation and maintenance of measure sets.

Internal development of measures to be used together to best capture quality for a certain purpose



External group creates a measure set to align measure use or for use in a certain system



© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

The following elements form a measure set: purpose, context, measure selection, data, implementation, and maintenance/feedback. A description of each element in the context of this work appears below.

PURPOSE: the aspect of quality that a set is measuring.

The first step in developing a measure set is to clearly identify the purpose, which informs all other aspects of measure set creation. The purpose establishes what aspect of quality the set is designed to measure and is related to a "quality construct." Previous NQF work on **composite measures,** which defines a quality construct as a "concept of quality" and identifies methods to construct composites, can inform development of the purpose for a measure set. The statement of purpose may be brief since related information is addressed more specifically in the context and measure selection sections, but it should be as specific as possible. For example, the measure set used in the CMS Hospital Star Ratings is designed to assess overall hospital quality,[21] and the purpose of the Joint Commission Hospital-Based Inpatient Psychiatric Services (HBIPS) measure set is to help hospitals compare their performance within hospital-based psychiatric services to that of their peers.[22]

CONTEXT: background details such as topic area, accountable entity, target population, setting, user(s), and intended use(s).

The purpose and the context work hand in hand. Both should be clarified upfront in measure set creation and clear to those being measured and those using performance results. Illustratively, the context of the measure set should consider the following:

- Measure sets can be created to focus on various topic areas including clinical areas such as cardiology or orthopedics or cross-cutting areas such as patient safety or primary care. Measure sets may also be categorized, for example, to address prevention, maintenance, or acute presentations.
- The accountable entity's (e.g., health plan, clinician, facility) performance is assessed by the measure set, which includes measures at various levels of analysis. This information should be transparent and aligned with the measure set's goal. Measures that have not been tested at the level in which they are intended to be used may need to go back through the NQF endorsement process to ensure the scientific merits of the measure are retained when modifications in specifications are made. Developers should consider how measure sets account for shared accountability when care coordination is required, but only one entity in the patient care continuum is held accountable (e.g., 30-day readmissions-only applying to the discharging entity and not including the adequacy of the receiving network of providers, nor their quality and efficiency). Measure sets may present the opportunity for a more coordinated approach to assess quality for complex measurement areas or to address measurement challenges like low case volume, incomplete or missing data, and attribution.
- The setting includes the location in which care is provided or the outcome is expected to occur.
- Population refers to the target population or the denominator population.
- Intended use refers to the model or way in which the set is to be used (e.g., for accountability applications like value-based purchasing and public reporting or for quality improvement). Specifying how measures sets should and should not be used can help prevent unintended consequences.
- "Users" refers to those implementing the set as well as those using the measure results to inform decision making or quality improvement (e.g., consumers, measured entities).

The defining components should be transparent and clearly communicated to users. If there is more than one intended use, each use should be delineated.

In addition to being transparent, these elements should be cohesive and consistent with the measure set purpose. It is likely not possible to set broadly applicable rules about the necessary consistencies of a measure set (e.g., to require that the level of analysis or setting be consistent for all measures within a measure set). If a developer seeks to align measures used to support a certain clinical topic or payment

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

model, design decisions may prioritize meeting this purpose over ensuring consistent target populations across all measures. Based on its purpose, a measure set may be a mixture of inpatient and outpatient indicators. It may be acceptable to have varying accountable units (e.g., hospital versus outpatient providers) if the measure set intends to capture the full care pathway of the target population. Measure sets may be limited by the measures available at the time of creation; when limitations exist in a measure set's specifications, a rationale should be provided.

MEASURE SELECTION: the process of choosing measures, the measures themselves, and how they fit a particular purpose.

Measure selection criteria likely differ depending on the purpose and context of a measure set. To determine whether the measures together meet a set's purpose, one can a) assess if the conceptual model and evidence base used to develop the measure set has face validity and is true to a set's purpose and b) assess if the measures together promote a parsimonious picture of performance for a topic area based on the purpose. For measure sets, qualitative methods (e.g., multistakeholder consensus) are often used to gauge whether the measures reflect the purpose.

Selection Principles

To create measure sets developers likely use selection principles to inform measure selection and removal. As part of this work, two national measure set selection principles used by the **Measure Applications Partnership (MAP)**[23] and the **Core Quality Measures Collaborative (CQMC)**[12] were considered. While it is not possible to have general selection principles that can be universally applied to all scenarios, these principles are a useful starting point to inform measure set design. These principles serve as core elements for measure set development, but measure selection principles may need to vary to support different, innovative use cases.

Building upon these selection principles, several points should be emphasized:

The measure selection principles and all of the measures should align with the purpose.

Measure set development should actively prioritize optimal alignment and reduced measurement burden.

- Effort toward harmonization should be included in the measure selection process as parsimonious, aligned measure sets can reduce burden.
- Measure sets that include measures that are not NQF-endorsed invite the potential for variability in validity, reliability, and feasibility. However, depending on a measure set's purpose, it may be more or less critical that the measures are NQF-endorsed. While NQF endorsement is the gold standard for accountability purposes, non-endorsed measures may play an important role in a set that serves another purpose.
- Outcome measures, or process measures closely linked to outcomes, should be prioritized.
- Key risk adjustment factors should be clear.
- Cost and efficiency measures may be used in certain measure sets, but their inclusion may have unique considerations.

All measures used in a measure set and their specifications should be available. Measure specifications that are readily available for implementors may help prevent problems due to variation in measure specifications and the use of outdated measure specifications. The measures' specifications should be tested and reviewed in a way that aligns with the measure set purpose. Evidence of the measures' reliability, validity, feasibility, and usability for the accountable unit should be transparent. NQF-endorsed measures have no need to repeat evaluation at the individual measure and measure set-level to prove they pass these criteria, as long as the context of the measure set and the level of analysis at which the measures were tested are aligned. Independently endorsed measures, however, may not always work together based on differences in relevant populations or exclusions that were not tested together.

Approach to Measure Set Use

There was limited agreement about whether a measure set must be implemented as a whole or if accountable units or implementers should have the ability to select measures from a measure set

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

to report on, referred to as a "pick list approach." However, most TEP members generally favored comprehensive measure sets without pick lists. Sets designed as an interlocking group of measures to be used in their entirety allow for greater completeness and parsimony. While a pick list approaches offers flexibility, it may create scores that do not truly represent a complete picture of quality and what unintended consequences can accrue if only partial sets are used. Using a pick list approach impacts the ability to meaningfully utilize benchmarks and compare performance between providers. For these reasons, NQF would not consider a pick list a measure set. It would not be appropriate for it to be assessed using the proposed evaluation method detailed later in this report as it does not operate as a true measure set.

Measure sets should be created so that all measures are used together to comprehensively assess the goal of the set.

While measure sets are increasingly being created as a method to promote alignment, reporting and collecting certain data remain challenges for some providers. This variability is one reason developers may allow a pick list option. Organizations may also feel that measure sets do not fully align with the care they provide or prefer using a measures set as a starting point for their value-based negotiations. An option that may be considered is to have a cross-cutting measure set with subsets underneath to allow for some flexibility to measure what is most applicable to a specific practice. However, this design option may still result in lack of alignment. Even with the ability to select certain measures to report, extensive measure lists contribute to measure burden. The best investment is likely in parsimonious sets that reflect the quality construct.

Certain scenarios need to be accounted for if measure sets are to be used in their entirety, for example, some providers may only able to report a portion of the required measures (e.g., due to minimum case requirements). Sample size considerations and patient mix need to be considered in development. Developers and stewards should also consider data issues that can arise from less flexibility. For some systems, one approach is to require data reporting from all entities, but those that do not meet pre-specified requirements would be left out of decisions or statements about quality.

Requiring that all measures be reported can lead to fewer entities receiving a score unless the majority of those reporting have the ability to meet a measurement system's reporting requirements. Details about how missing data are handled and the implication to overall performance inferences should be transparent. Various statistical methods or imputation logic may be used to overcome missing or incomplete data. If these methods are used, they should support the goal of the measurement system and be accompanied by a reasonable rationale. Considerations around incomplete and missing data require separate, rationalized approaches, but should not prevent the development of comprehensive measure sets. Methods used to infer performance should be used if it is not possible to report. It should not be a tool used by those who simply prefer not to report, a distinction which may be difficult to discern.

When using measure sets, there is a need to consider if providers can report all measures and how missing data will be handled.

Unintended Consequences

Several issues and solutions were discussed related to potential negative unintended consequences of measure sets. If measures are tested and endorsed separately and not evaluated together, it may be difficult to track and understand the consequences of measurement until several years post implementation. For example, measuring hospital length of stay, use of emergency department observation stays, and readmissions or pain management and opioid overutilization separately, it may not be possible to identify unintended consequences to patients. The development of a measure set can help bring together balancing measures to help prevent unintended consequences. Balancing measures consider a system from different dimensions to ensure improvement in one part of the system is not creating problems in other parts.[24] Developers should also examine any evidence that measures in a set are highly performing with little room for additional improvement or variation (e.g.,

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

topped out, ceiling effect). If measures are topped out, developers should provide justification to include them. One rationale for their inclusion may be, for example, to gauge performance across a larger population than those currently reporting or to assess performance in a population with known disparities.

DATA: the information sources and collection methods.

Developers should provide information about how measure data will be obtained (e.g., abstracted from patient charts or using claims data). When a measure set is deployed, the data collection and reporting approach should be consistent among all entities included in a benchmark. For example, benchmarks that use data from eCQMs versus claims-based reporting may need to be different. If there are measure-specific, minimum sample size recommendations, they should be transparent. Establishing consistent data collection methods and requirements promotes consistent and appropriate use of the measure set as intended. Other characteristics of the measure set (e.g., context, timing, reliability, accountable entity, and target population) are related to data decisions and need to be taken into account.

Feasibility of data collection should be considered in measure set development. Interoperability of data for reporting and the ability to access and use the results should be considered. Measurement is increasingly seeking opportunities to move toward more standardized specifications and less burdensome reporting. To this end, measures with these characteristics, like eCQMs, may be prioritized for inclusion in measure sets.

A measure set should produce data that are actionable. There is a need to get data to end users in a more timely manner. End users may be consumers, the entity whose performance is being assessed, or the entity who is using the results to make decisions.

IMPLEMENTATION: guidance provided to users about how the measure set is to be implemented.

Implementation can be thought of in two ways—implementation by an entity that has to submit data and implementation by a payer or other group assessing the performance of the reporting entity. Guidance regarding implementation should address both audiences. Without clear guidance about how a measure set should be executed, users may implement measures in a way that is not supported by their testing or use the measure set in a way that is not true to its purpose. Users may also modify measure set specifications; this variation may limit the ability to make fair comparisons. For these reasons, the purpose and intended use of the measure set should be clearly communicated as part of implementation guidance.

Guidance should be complete, precise, and prescriptive to ensure each entity is using the same procedures and processes. Those being measured and those using the results should understand the types of measures, how they are to be reported, what the data needs are, and how the measures are related. The measures within a measure set need to be clearly defined and articulated, including information on the numerator, denominator, exclusions, population, setting, level of analysis, and risk adjustment (if applicable). Other useful information that should be considered include an FAQ document, examples of correct and incorrect specifications and reporting, and training for data abstraction. Developers may also consider including the frequency at which performance should be evaluated using the measure set.

MAINTENANCE/FEEDBACK: processes for updating the measure set and communicating performance results.

Maintenance refers to the use of a process to ensure a measure set remains updated, for example, with changes in evidence or performance. The developer of a measure set should periodically review the evidence supporting the measure set, the measure set's impact, and whether there is still an opportunity for performance improvement. The measure set developer should determine the cadence for maintenance review. If individual measures are up for

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

NQF endorsement maintenance, then the measure set developer or steward should review the final decision and rationale to determine if a measure endorsement status change impacts the inclusion of a measure in the set. Substantive measure changes should trigger a review of a measure set.

There should be an early opportunity after introduction of the measure set to hear from those being measured about their experience with the measure set. Feedback from those being measured and those using the measures should be considered in the measure set maintenance process. Measure sets should be open for public comment when possible to capture opportunities to advance them. Stakeholders should have the opportunity to comment on the use of the measure set, the feasibility of its use within their systems, perceived barriers, and unintended consequences.

Several examples were provided to illustrate ways to determine during maintenance if measures sets are serving the purpose of their design:

- Movement toward population health goals, positive change in desired outcomes, improved efficiency, and enhanced value to the impacted patient population and the healthcare system
- Reviewing how widespread use of the measure set is
- Ensuring the measures within the set have sufficient denominator size or sample size for validation and use
- Ability to determine high-and-low performance outliers
- Evaluating the data capture burden or feasibility and usability
- Evaluating improvement in performance either on individual measures or at an aggregated level
- Utilization of the measure set in benchmarking and ability to use the set to identify meaningful differences between entities
- Rapidity with sharing results back to care providers
- Changes in the processes of the reporting entities, policy, or payment structure

Feedback refers to the communication of measure results so they can be used to drive performance improvement. Stakeholders impacted by measure set use and those who would benefit from understanding the results (e.g., entities reporting the measure, consumers using the service for which the measures reflect quality, payers, regulators, and policymakers) should reasonably have timely access to performance results. While more frequent data, benchmarking updates, and shorter look-back periods may inform more real-time improvement, these decisions must be balanced with having enough data to ensure performance results are reliable and valid. A measure set developer may not be positioned to know how those being measured will access measure results. Sharing performance data may be a responsibility of the end user of the measure set (e.g., implementer, rater/ranker).

MEASUREMENT SYSTEM DESIGN ELEMENTS

A measurement system is a group of measures that, based on a predefined methodology, work together to assess quality or cost in relationship to a goal. While several elements of measure sets are similar to measurement system elements, there are also important distinctions. Furthermore, a pre-established method to determine performance of entities relative to one another is a distinct aspect of a measurement system. Measurement systems contain a measure set plus other components. The elements that make up a measurement system include goal, context, measure selection, measure grouping, scoring approaches, risk adjustment, and usability. Each element and the considerations that are unique to measurement systems are described below.

The elements of a measurement system should be transparent and clearly communicated.

GOAL: the objective that the system is assessing.

Defining a clear goal of the measurement system is central to its design and evaluation. A clear goal is important to guide the approach and understand trade-offs with measure selection, grouping, and scoring. For example, the goal of a hospital rating

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

program can be characterized as a consumer-facing tool that allows a comprehensive assessment of the quality of care provided by US hospitals. This goal would have implications to the types of measures, domains, and weighting design decisions to ensure generalizability while recognizing the range of clinical services offered. Special attention should be given to each element of a measurement system design to ensure that it ultimately aligns with the stated goal.

System design should start with the conceptualization of the specific goal and consider unintended consequences.

There should be principles set during the development of a measurement system. These principles help establish what a measurement system is trying to achieve and should be clear and transparent to stakeholders whom the system is intended to inform. For example, principles might include making performance information easy to understand, useful for patients and consumers, scientifically valid, generalizable and inclusive of all accountable units, and transparent of methodological decisions or trade-offs. Developers should also explicitly take into account how the measurement system relates to other existing ones.

There is a need to promote the efficient use of measurement resources through system design and to do so without limiting the creation of innovative or unique systems to suit specific needs.

CONTEXT: background details such as accountable entity, intended use, incentive structure, measurement periodicity, and attribution method.

Measurement systems can be designed to assess performance of hospitals, clinicians, or health plans. These systems should be transparent on how care and services are specifically attributed to the measured entity and include a rationale on how the accountable unit can reasonably influence the outcomes measured in the system. Further, the intended audience of the measurement system might include consumers, health plans, or government agencies. These various accountable units and intended users may necessitate different design decisions of the underlying measures used. Contextual considerations should also include whether the measurement system is voluntary or requires participation. Mandatory participation in measurement systems may need to account for the range of measures and sample size requirements to facilitate participation.

The incentive structure in which the measurement system is deployed should also be considered. Broadly, it is important to understand if the measurement system is intended for public reporting or payment applications. However, it is important to identify specifically how the measures will be deployed. For example, does the measurement system use a classification system? Is the performance score point estimate used in the measurement system? Or does the system specifically account for measurement error in the performance score? It is also important to consider how specifically the measurement system translates to financial rewards or penalties for the accountable units. For example, are bonuses given to top-ranked providers or are bonuses distributed based on point estimate thresholds of performance? Measurement systems should also consider how measure performance may impact reimbursement differentially over time. For example, for MIPS, 4% of provider reimbursement was at risk in 2019, rising to 9% of payment at risk in 2022.[25] Taken together, these contextual factors are critical to the design and evaluation of measurement systems.

MEASURE SELECTION: the process of choosing and retiring measures, the measures themselves, and how they reflect the goal.

The selection of specific performance measures is a central building block to both measure sets and measurement systems. Fundamentally, individual performance measures should meet scientific standards for evidence, reliability, validity, feasibility, and usability. However, in the deployment of new or innovative performance measures, some flexibility in these standards may be required while evidence is being generated.

When considering scientifically sound measures for application in a measurement system, additional factors should be considered. The inclusion and

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

exclusion criteria for measure selection should be transparent and aligned with the measurement system goal and context in which they are deployed. Measures used in a measurement system should align to the extent possible the measurement time period to ensure assessments of provider performance capture a similar timeframe and number of years of data required to achieve reliable measure score results. The system should account for small numbers in terms of ability to report or sample size in number of patients within the accountable unit. Exclusion criteria should consider non-directional measures in which it is not clear whether higher or lower score performance is better, overlapping measures in cohort and/or outcome. All final measure specifications should be transparent and publicly available.

MEASURE GROUPING: how measures are aggregated or assigned to domains.

Measurement systems may deploy the use of measure groupings or domains to group individual measures. These groupings may be normative based on clinical coherence, a policy objective, or principles identified in the goal of the measurement system. For example, normative groups may include clinical areas, such as cardiovascular or cancer care, or patient experience and cost and efficiency. The groups may also be derived empirically using statistical methods that take into account the psychometric properties of the measures used in the system. For example, do the measures load on expected factors and how correlated is measure performance across measures in the system? Either or both methods may be appropriate given the goal and context of the measurement system. Public sector programs that require mandatory reporting, in particular, should allow multistakeholder input on the design of the measure grouping approach. Private sector or voluntary programs should make the rationale and decisions on measure grouping transparent.

SCORING APPROACHES: the methods by which overall performance is determined.

The development of measure systems requires the aggregation of individual performances into an aggregate score. There are a range of scoring methods that can be used depending on the goal and the context of the measurement system. For example, the use of a latent variable modeling (LVM) approach can be used to estimate a group score for multiple measure groups. This approach uses a statistical model that assumes multiple measures reflect a single unobserved latent quality trait of an accountable unit that cannot be directly measured.

System design decisions are often value judgements. Methods should be transparent, statistically appropriate, and aligned across programs when possible.

The LVM approach attempts to measure this underlying quality trait through correlation and variation of measures in a given group.[26] In the design of measurement systems, complexity of modeling decisions should also be balanced with intuitiveness and clarity especially for those being measured. The methods of standardizing measure scales (e.g. use of z-scores), weighting, and assignment of summary scores to categories of performance should reflect an underlying logic. For example, a K-means clustering method can be used to assign summary scores to a five Star Rating methodology. This approach separates scores into categories such that hospital summary scores are most like scores in the same category and least like summary scores in other categories.[27] The method for K-means clustering can be used to derive statistically based cut points. These cut points can be used to establish thresholds of performance for each star change across time periods based on the underlying measures and performance distributions across accountable units. The approach to weighting individual measures or measure groups should be transparent, recognizing that the approach to weighting may be a value judgement based on the goals and context of the measurement system. Assignment of summary scores to performance categories can be done using predefined benchmarks, policy-driven objectives, or empirically derived methods. The impact of weighting of individual measures or measure groups on the scoring or ranking of measured entities should be clear. The scoring approach should also make transparent requirements of minimum sample size for reporting and how missing data is handled and not create any inherit bias in the results.

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

RISK ADJUSTMENT: the approach to isolate quality differences by accounting for differences in patient mix across entities.

Risk adjustment can occur at multiple levels of a measurement system. Individual performance measures can be risk adjusted to reflect differences in patient mix, clinical severity, or social risk. Other forms of standardization may be deployed depending on the measure type. For example, cost and efficiency measures may deploy a price standardization approach to account for geographic differences in input costs. Additionally, measurement systems can create peer groups or other methods of stratification based on patient risk factors. Regardless of whether a system risk adjusts individual measures or uses standardization or stratification, these approaches should be transparent and include the variables used to adjust and the impact of the adjustment. These methods should be consistent with the goals and context of the measurement system.

USABILITY: how the methods and performance results are communicated.

A measurement system should be transparent with the methodologies deployed. Specifically, the approach to how performance information will be deployed for the intended user. Consideration should be made to make de-identified data and coding files available to the public. Providers should be aware of measures they are being held accountable for before being receiving performance results.

It is necessary to ensure usability of the system and actionability of the results by relevant audiences, especially consumers.

Recommendations for Evaluation

ROLE OF A CONSENSUS-BASED ENTITY

NQF recognizes there are links between this work and the various other NQF programs, including measure endorsement and measure selection by the MAP. There may be the need to emphasize certain areas during individual measure endorsement: interaction of a measure with other measures, additional analysis of real-world unintended consequences or benefits, and in which sets or systems a measure would be appropriate for inclusion. The framework described in this report can advance the measurement enterprise toward a more comprehensive approach to healthcare quality assessment. Furthermore, it supports the need for holistic review not only of the measures that should be added to programs, but more importantly, of how a measurement system comprehensively functions to achieve its goal. For example, NQF demonstrated the benefit of this type of multistakeholder review by leading the Hospital Quality Star Ratings Summit, which produced practical recommendations and feedback toward program improvement.

Advancing health outcomes for all remains paramount. Consumers are seeking greater engagement in their healthcare decisions. Payment is increasingly being tied to quality performance. We must ensure the quality measurement infrastructure is optimally designed to support these needs. As NQF explores opportunities to operationalize standardized, consensus-based review of measure sets and measurement systems, we will pay close attention to employ a streamlined approach that ensures measurement drives health improvement and accurately identifies and rewards the delivery of high-quality care to all patients. This is NQF's first effort to establish a standardized approach for the review of measure sets and measurement systems. NQF looks forward to working closely with the private and public sectors in achieving this goal.

APPROACH TO EVALUATION

NQF worked with TEP to develop an approach to evaluation by identifying key design elements for measure sets and measurement systems. These elements are translated into submission forms for measure sets and measurement systems as a tool for evaluation. The submission forms, which can be found in Appendix B: Measure Set Submission Forms and Appendix C: Measurement System Submission

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Form, outline the information that would be required to be submitted and evaluated by a multistakeholder committee to assess a measure set or measurement system. Making this information transparent is a crucial first step in a standardized evaluation of measure sets and measurement systems, and the feedback that would be elicited though this consensus-based process would inform areas for improvement for a measure set or measurement system. Each submission form section centers around an element of a measure set or measurement system and includes questions to obtain the necessary information about its subcomponents.

The rationale for measure set and measurement system design decisions should be subject to multistakeholder review.

Appendix D provides a side-by-side comparison of the measure sets and measurement systems submission forms to display the relationship among their respective elements. While there is overlap between measure sets and measurement systems, the elements of each include distinct content. For example, both sets and systems include the element "Context." However, "Context" for measurement systems includes unique information that is not relevant to measure sets, such as the incentives or disincentives of the system, the frequency at which the system evaluates performance, and the attribution method. Further, Appendix D provides additional clarification of submission form elements unique to measurement systems, specifically measure grouping, scoring approaches, and risk adjustment.

The responses provided would be reviewed by NQF and discussed by a multistakeholder committee. If developers or stewards were unable to provide responses to the questions, a rationale would be required. A multistakeholder review would involve an assessment, including discussion and feedback, of each element in order, starting with the purpose or goal and concluding with an overall assessment. Important to emphasize in this process is the need for a measure set or measurement system to be evaluated based on its intent. For example, if a measure set is intended to be used for quality improvement versus payment adjustment, the design decision would be discussed in this context. The best case scenario, however, may be a measure set that is useful for impacting quality improvement and making payment adjustments.

Evaluation should consider the transparency and appropriateness of the decisions made in set or system design based on the specified purpose or goal.

EVALUATION CONSIDERATIONS

Since there are overlapping elements between measure sets and measurement systems, both review processes would be structured similarly. As a point of distinction, the measure set review process would be appropriate for measure sets that are designed for use in multiple systems. Such a measure set would be reviewed agnostic from its use in one particular system, and assessment would serve to verify that the measure set's design is appropriately aligned with its purpose.

Flexibility in these proposed evaluation methods may be necessary. NQF recognizes that complete transparency may be problematic for measurement systems that employ propriety elements, for example, value-based payment models used by the private sector.

Similar to the NQF's individual measure endorsement process, the process for reviewing sets and systems is expected to evolve over time as measurement science advances.

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Conclusion

To build upon the use of individual measures to assess quality, safety, and person-centered care, NQF has created a standardized, transparent method to define and assess the design of measure sets and measurement systems. Placing emphasis on intent at the forefront, ensuring the methodology is aligned, and involving diverse stakeholders in the discussion are essential to ensuring measurement systems provide accurate inferences about quality of care.

There exists an opportunity in the healthcare quality measurement field for greater access to comprehensive, actionable, and scientifically sound data to compare healthcare quality. Understanding the importance of the relationship between observed performance outcomes and the design of measurement systems is essential to the evolution of healthcare quality measurement and value-based models. As a leader in the field, NQF recognizes the challenge and importance of proactively aligning measures across stakeholders and improved transparency. NQF welcomes passionate partner organizations in the uphill transformation necessary to move healthcare measurement forward in the most efficient and effective way. Advancing the current state of quality will take every voice driving the field toward higher levels of performance.

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

References

1 Institute of Medicine (US) Committee on Quality of Health Care in America. To *Err Is Human: Building a Safer Health System.* (Kohn LT, Corrigan JM, Donaldson MS, eds.). Washington (DC): National Academies Press (US); 2000. http://www.ncbi.nlm.nih.gov/books/NBK225182/. Last accessed July 2020.

2 Institute of Medicine (US) Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington (DC): National Academies Press (US); 2001. http://www.ncbi.nlm.nih.gov/books/NBK222274/. Last accessed July 2020.

3 Rechel B, McKee M, Haas M, et al. Public reporting on quality, waiting times and patient experience in 11 high-income countries. *Health Policy*. 2016;120(4):377-383.

4 Campanella P, Vukovic V, Parente P, et al. The impact of Public Reporting on clinical outcomes: a systematic review and meta-analysis. *BMC Health Serv Res.* 2016;16. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4957420/. Last accessed May 2020.

5 Center for Medicare and Medicaid Services. Hospital Compare overall rating. Data.Medicare.Gov. https://data.medicare.gov/data/hospital-compare. Published 2018. Last accessed May 2020.

6 Bilimoria KY, Birkmeyer JD, Burstin H, et al. Rating the Raters: An Evaluation of Publicly Reported Hospital Quality Rating Systems. *NEJM Catal.* August 2019. https://catalyst.nejm.org/doi/abs/10.1056/CAT.19.0629. Last accessed May 2020.

7 Gilstrap L, Skinner J, Barbara G, et al. Opportunities and Challenges of Claims-Based Quality Assessment: The Case of Postdischarge β-Blocker Treatment in Patients With Heart Failure With Reduced Ejection Fraction. *Circ Cardiovasc Qual Outcomes*. 2020;13(3):e006180.

8 Landrum MB, Nguyen C, O'Rourke E, et al. *Measurement Systems: A Framework for Next Generation Measurement of Quality in Healthcare*. Washington, DC: National Quality Forum; 2019.

9 National Quality Forum. NQF: National Quality Forum Hospital Quality Star Rating Summit. http://www.qualityforum.org/NQF_Hospital_Quality_Star_Rating_Summit.aspx. Last accessed May 2020.

10 National Quality Forum. *Composite Performance Measure Evaluation Guidance*. Washington, DC: National Quality Forum; 2013. https://www.qualityforum.org/Publications/2013/04/Composite_Measure_Guidance_Final_Report.aspx.

11 National Quality Forum. NQF: CQMC Core Sets. https://www.qualityforum.org/CQMC_Core_Sets.aspx. Last accessed May 2020.

12 National Quality Forum. CQMC: Core Quality Measures Collaborative. http://www.qualityforum.org/cqmc/. Last accessed May 2020.

13 National Quality Forum. *A Core Set of Rural-Relevant Measures and Measuring and Improving Access to Care: 2018 Recommendations from the MAP Rural Health Workgroup.* Washington, DC: National Quality Forum; 2018. http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=88226.

14 Agency for Healthcare Research and Quality. Major Hospital Quality Measurement Sets. http://www.ahrq.gov/talkingquality/measures/setting/hospitals/measurement-sets.html. Published May 2019. Last accessed May 2020.

15 The Joint Commission. What you need to know: Frequently Asked Questions (FAQs) 2020 ORYX® Performance Measure Reporting Requirements. https://www.jointcommission.org/-/media/tjc/documents/measurement/oryx/cy2020_what_you_need_to_know-oryx_faqs.pdf.

16 Center for Medicare and Medicaid Services. How are hospital overall ratings calculated? Hospital Compare. https://www.medicare.gov/hospitalcompare/Data/Hospital-overall-ratings-calculation.html. Last accessed May 2020.The Joint Commission. What you need to know: Frequently Asked Questions (FAQs) 2020 ORYX® Performance Measure Reporting Requirements. https://www.jointcommission.org/-/media/tjc/documents/measurement/oryx/cy2020_what_you_need_to_know-oryx_faqs.pdf.

17 Center for Medicare and Medicaid Services. Quality Payment Program: Merit-based Incentive Payment System (MIPS) Overview. https://qpp.cms.gov/mips/overview. Last accessed May 2020.

18 Center for Medicare and Medicaid Services. Shared Savings Program: About The Program. CMS.gov. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/about. Last accessed May 2020.

19 Center for Medicare and Medicaid Services. Medicare Shared Savings Program Quality Measure Benchmarks for the 2019 Performance Year. February 2020. https://www.cms.gov/files/document/2019-ssp-quality-measures-benchmark.pdf.

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

20 Center for Medicare and Medicaid Services. Medicare Shared Savings Program Quality Measurement Methodology and Resources Specifications. May 2019. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/quality-measurement-methodology-and-resources.pdf.

21 Center for Medicare and Medicaid Services. Overall hospital quality star rating. Hospital Compare. https://www.medicare.gov/hospitalcompare/About/Hospital-overall-ratings.html. Last accessed May 2020.

22 National Association for Behavioral Healthcare. HBIPS Core Measures. Hospital-Based Inpatient Psychiatric Services Core Measure Set. https://www.nabh.org/policy-issues/quality/hbips-core-measures/. Last accessed May 2020

23 National Quality Forum. Measure Applications Partnership: MAP Member Guidebook. November 2019. http://www.qualityforum.org/Projects/i-m/MAP/MAP_Member_Guidebook.aspx.

24 Institute for Healthcare Improvement. Science of Improvement: Establishing Measures. http://www.ihi.org/resources/Pages/HowtoImprove/ScienceofImprovementEstablishingMeasures.aspx. Last accessed May 2020.

25 Center for Medicare and Medicaid Services. Quality Payment Program: Merit-Based Incentive Payment System (MIPS) 101 Guide, 2019 Performance Year. April 2020. https://qpp-cm-prod-content.s3.amazonaws.com/uploads/607/2019%20MIPS%20101%20Guide.pdf.

26 Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modeling. *Psychometrika*. 2004;69(2):167-190.

27 MacQueen J. Some methods for classification and analysis of multivariate observations. *In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, CA: University of California Press; 1967:281-297. https://projecteuclid.org/

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Appendix A: Technical Expert Panel Rosters and NQF Staff

TECHNICAL EXPERT PANEL CO-CHAIRS

MEASURE SETS CO-CHAIR

Amy Nguyen Howell, MD, MBA, FAAFP
Chief Medical Officer, America's Physician Groups
Los Angeles, California

MEASUREMENT SYSTEMS CO-CHAIR

Michael Chernew, PhD
Harvard Medical School
Boston, Massachusetts

MEASUREMENT SYSTEMS CO-CHAIR

Sam Simon, PhD
Associate Director, Mathematica
Cambridge, Massachusetts

MEASUREMENT SETS SUBGROUP

Thomas Aloia, MD
Chief Value and Quality Officer, University of Texas, MD Anderson Cancer Center
Houston, Texas

Tricia Elliott, MBA, CPHQ
Director Quality Measurement, The Joint Commission
Oakbrook Terrace, Illinois

Lindsay Erickson
Director, Program Operations, Integrated Healthcare Association
Oakland, California

Louis Galterio, MBA
President, Suncoast RHIO
North Port, Florida

Frank A. Ghinassi, PhD, ABPP
President and CEO, Rutgers Health University, Behavioral Health Care, Rutgers Health
Piscataway, New Jersey

Denise Morse, MBA
Director, Quality and Value Analytics, City of Hope
Duarte, California

Matthew Pickering, PharmD
Senior Director, Research & Quality Strategies, Pharmacy Quality Alliance
Alexandria, Virginia

E. Clarke Ross, DPA
Public Policy Director, American Association on Health and Disability
Rockville, Maryland

Mathew Sapiano, PhD
Statistician, Centers for Disease Control and Prevention
Atlanta, Georgia

Sharon Sutherland, MD, MPH
Physician, Cleveland Clinic
Cleveland, Ohio

Traci Thompson Ferguson, MD, MBA, CPE
Chief Medical Director, Medical Management and External Relationships, WellCare Health Plans, Inc.
Tampa, Florida

MEASUREMENT SYSTEMS SUBGROUP

Philip M. Alberti, PhD
Senior Director, Health Equity Research and Policy, Association of American Medical Colleges
Washington, District of Columbia

J. Matthew Austin, PhD
Faculty, Johns Hopkins School of Medicine
Baltimore, Maryland

Kari Baldonado
Senior Director and Solution Executive, Quality Measurement, Cerner Corporation
Kansas City, Missouri

Julie Bershadsky, PhD
Senior Research Associate, Human Services Research Institute
Cambridge, Massachusetts

Amy Chin, MS
Senior Director, Health Economics and Outcomes Research, Greater New York Hospital Association
New York, New York

William Conway, MD
Chief Executive Officer, Henry Ford Medical Group
Detroit, Michigan

Missy Danforth
Vice President, Hospital Ratings, The Leapfrog Group
Washington, District of Columbia

Marybeth Farquhar, PhD, MSN, RN
Executive Vice President, Research, Quality & Scientific Affairs, American Urological Association
Linthicum, Maryland

Danielle Lloyd, MPH
Senior Vice President, Private Market Innovations & Quality Initiatives, America's Health Insurance Plans
Washington, District of Columbia

Jeffrey Sussman, PhD, MPH
Senior Associate, Booz Allen Hamilton
Washington, District of Columbia

FEDERAL GOVERNMENT MEMBER

Michelle Schreiber, MD
Director, Quality Measurement & Value Based Incentives Group, CMS
Baltimore, Maryland

NQF STAFF

Sheri Winsper, RN, MSN, MSHA
Senior Vice President, Quality Measurement

Maha Taylor, MHA, PMP
Managing Director, Quality Measurement

Nicolette Mehas, PharmD
Director

Teresa Brown, MHA, MA, CPHQ, CPPS
Senior Manager

Madison Jung
Manager

Yvonne Kalumo-Banda, MS
Manager

Amy Guo, MS
Analyst

Asaba Mbenwoh Nguafor, RN, MSN/MPH
Analyst

Taroon Amin, PhD, MPH
Consultant

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Appendix B: Measure Set Submission Forms

Element of a measure set	Developer/Steward Submission Items	Decision Categories
Purpose	• How do you define the quality construct of the measure set (i.e., aspect of quality the set is measuring (e.g., overall hospital quality, quality of inpatient psychiatric care))?	N/A
Context	• Describe the following components of the set: topic area, accountable entity, target population, and setting. • Are these components consistent across all measures in the set? • Is there appropriate consistency across measures relative to the intended use of the set? • What is the intended use(s) (e.g., value-based payment, public reporting, quality improvement)? • Who are the intended users(s) (e.g., payers, hospitals)?	Support; Conditional support; Do not support
Measure Selection	• What are the criteria or principles used for measure inclusion or exclusion? • Please list the measures and data sources included in the set with references to specifications. What is the evidence or rationale to support the use of the measures? • For each measure, summarize evidence, reliability, validity, feasibility, and usability information for the accountable unit. • Is there evidence any of the measures are topped out? If so, what is the rationale for retaining the measure? • Do the measures in the set accurately reflect the intended purpose? How?	Support; Conditional support; Do not support
Data	• What is the data source? • How will data be collected and verified? Is it summary or person-level? • How are missing data handled? • What is the time frame for data collection? • What is the minimum sample size for reporting?	Support; Conditional support; Do not support
Implementation	• Are all measures in the set designed to be used together? • What guidance is provided to users of the measure set, especially if all measures in the set are not designed to be used together?	Support; Conditional support; Do not support
Maintenance and Feedback	• What are the processes to ensure the measure set remains updated (e.g., what are the sources of information to update measures? What is the frequency of measure updates?)? • What is the feedback process by which those being measured have access to measure results? • How and when is success of the set determined (e.g., widespread use of the set, improvements in performance scores)?	Support; Conditional support; Do not support
Overall	• Is the measure selection approach, information about data, implementation process, and maintenance/feedback process consistent with the construct of the measure set?	Y-Support N-Do not support

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Appendix C: Measurement System Submission Form

Element of a measurement system	Developer/Steward Submission Items	Decision Categories
Goal	• What is the objective of the measurement system (e.g., what is the system trying to assess?)? • What principles were used in developing the measurement system? • Which stakeholder groups are affected by this objective? • How were stakeholder groups involved in the development? • Provide a rationale for the goal of the system. Is there overlap with any other existing systems?	N/A
Context	• What is the accountable unit of the measurement system (e.g., physician, hospital)? • How is the measurement system intended to be used (e.g., public reporting, value-based payment)? • What is the incentive structure in place for the accountable units being measured (e.g., percent of dollars for performance above the mean)? • If the measurement system is intended for public reporting or accountability, please include discussion of possible incentives or disincentives of the system. • How often is performance evaluated using the system? • Are the performance periods for measures aligned? • What is the attribution model/method of the system (e.g., how is care attributed to a measured entity?)? • How will the accountable entity influence the measures in the system? • Is the system voluntary or compulsory?	Support; Conditional support; Do not support
Measure Selection	• Describe the method of identifying measures for potential inclusion and method for measure removal. • What are the criteria used for measure inclusion or exclusion? • How is the measure selection criteria consistent with the goal of the measurement system? • List measures and data sources included in the measurement system with references to specifications. • What is the performance for each measure at the level of the measured entity? • Is there justification for retaining topped-out measures, if applicable? • For each measure, summarize the evidence, reliability, validity, feasibility, and usability information for the accountable unit. • How do the measures reflect the goal of the system?	Support; Conditional support; Do not support
Measure Grouping	• If the measurement system includes measure grouping or domains, please list the groups/domains and the measures in each. • What is the method of measure aggregation (normative—based on subject matter expertise or following certain principles; empirical—based on statistical methods?)? • Have the groups been tested for this purpose? • Is there a hierarchy? If so, what does it look like?	Support; Conditional support; Do not support

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Element of a measurement system	Developer/Steward Submission Items	Decision Categories
Scoring Approaches	• What statistical techniques are used as part of scoring? • Are cut points used for scoring? If so, how are they created? • Are reference and/or peer groups used in scoring? If so, what is the impact on scoring? • What methods are used for standardizing measure scales? • How reliable is the scoring method (e.g., reliability at various component levels of the hierarchy)? • Is improvement considered as part of scoring? • What is the potential for misclassification? • What are the rules with regard to weighting? • Please specify if methods are used on particular domains or overall, if applicable. • How will data be validated? • Are multiple years of data used in scoring? • Is data summary or person-level? • How are missing data handled? • What is the minimum sample size for reporting eligibility? • What is the minimum sample size for the denominator at the measure level? • How are the scoring approaches consistent with the goal of the measurement system?	Support; Conditional support; Do not support
Risk Adjustment	• What is the approach to risk adjustment, standardization, or stratification, if any? • For which variables would you adjust or stratify? • Is risk adjustment at the individual measure or aggregate level? • Please provide a conceptual rationale if risk adjustment or peer grouping is not part of the measurement system. • Is there a conceptual rationale for the impact of social factors on scoring? • Was social risk adjusted considered and tested? • How is the risk adjustment approach consistent with the goal of the measurement system?	Support; Conditional support; Do not support
Usability	• Are the methodologies used in the measurement system transparent? • Will the data be available? • How will performance information be displayed? • Who is the intended user of this information (e.g., consumers, health plans, government agency)?	N/A
Overall	• Is the approach to measure selection, measure grouping, scoring, and risk-adjustment consistent with the goal of the measurement system?	Y-Support N-Do not support

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Appendix D: Measure Sets and Measurement Systems Element Comparison

Measure Set Element and Description	Measurement System Element and Description
Purpose	**Goal**
1) How do you define the quality construct of the measure set (i.e., aspect of quality the set is measuring (e.g., overall hospital quality, quality of inpatient psychiatric care))?	1) What is the objective of the measurement system (e.g., what is the system trying to assess?)? 2) What principles were used in developing the measurement system? 3) Which stakeholder groups are affected by this objective? 4) How were stakeholder groups involved in the development? 5) Provide a rationale for the goal of the system. Is there overlap with any other existing systems?
Context	**Context**
1) Describe the following components of the set: topic area, accountable entity, target population, and setting. a) Are these components consistent across all measures in the set? b) Is there appropriate consistency across measures relative to the intended use of the set? 2) What is the intended use(s) (e.g., value-based payment, public reporting, quality improvement)? 3) Who are the intended users(s) (e.g., payers, hospitals)?	1) What is the accountable unit of the measurement system (e.g., physician, hospital)? 2) How is the measurement system intended to be used (e.g., public reporting, value-based payment)? 3) What is the incentive structure in place for the accountable units being measured (e.g., percent of dollars for performance above the mean)? 4) If the measurement system is intended for public reporting or accountability, please include discussion of possible incentives or disincentives of the system. 5) How often is performance evaluated using the system? 6) Are the performance periods for measures aligned? 7) What is the attribution model/method of the system (e.g., how is care attributed to a measured entity?)? 8) How will the accountable entity influence the measures in the system? 9) Is the system voluntary or compulsory?
Measure Selection	**Measure Selection**
1) What are the criteria or principles used for measure inclusion or exclusion? 2) Please list the measures and data sources included in the set with references to specifications. What is the evidence or rationale to support the use of the measures? 3) For each measure, summarize evidence, reliability, validity, feasibility, and usability information for the accountable unit. 4) Is there evidence any of the measures are topped out? If so, what is the rationale for retaining the measure? 5) Do the measures in the set accurately reflect the intended purpose? How?	1) Describe the method of identifying measures for potential inclusion and method for measure removal. 2) What are the criteria used for measure inclusion or exclusion? 3) How is the measure selection criteria consistent with the goal of the measurement system? 4) List measures and data sources included in the measurement system with references to specifications. 5) What is the performance for each measure at the level of the measured entity? 6) Is there justification for retaining topped-out measures, if applicable? 7) For each measure, summarize the evidence, reliability, validity, feasibility, and usability information for the accountable unit. 8) How do the measures reflect the goal of the system?

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Measure Set Element and Description	Measurement System Element and Description
Data	**Measure Grouping**
1) What is the data source? 2) How will data be collected and verified? Is it summary or person-level? 3) How are missing data handled? 4) What is the time frame for data collection? 5) What is the minimum sample size for reporting?	1) If the measurement system includes measure grouping or domains, please list the groups/domains and the measures in each. 2) What is the method of measure aggregation (normative—based on subject matter expertise or following certain principles; empirical—based on statistical methods?)? 3) Have the groups been tested for this purpose? 4) Is there a hierarchy? If so, what does it look like?
Implementation	**Scoring Approaches**
1) Are all measures in the set designed to be used together? 2) What guidance is provided to users of the measure set, especially if all measures in the set are not designed to be used together?	1) What statistical techniques are used as part of scoring? 2) Are cut points used for scoring? If so, how are they created? 3) Are reference and/or peer groups used in scoring? If so, what is the impact on scoring? 4) What methods are used for standardizing measure scales? 5) How reliable is the scoring method (e.g., reliability at various component levels of the hierarchy)? 6) Is improvement considered as part of scoring? 7) What is the potential for misclassification? 8) What are the rules with regard to weighting? 9) Please specify if methods are used on particular domains or overall, if applicable. 10) How will data be validated? 11) Are multiple years of data used in scoring? 12) Is data summary or person-level? 13) How are missing data handled? 14) What is the minimum sample size for reporting eligibility? 15) What is the minimum sample size for the denominator at the measure level? 16) How are the scoring approaches consistent with the goal of the measurement system?
Maintenance and Feedback	**Risk Adjustment**
1) What are the processes to ensure the measure set remains updated (e.g., what are the sources of information to update measures?, what is the frequency of measure updates?)? 2) What is the feedback process by which those being measured have access to measure results? 3) How and when is success of the set determined (e.g., widespread use of the set, improvements in performance scores)?	1) What is the approach to risk adjustment, standardization, or stratification, if any? 2) For which variables would you adjust or stratify? 3) Is risk adjustment at the individual measure or aggregate level? 4) Please provide a conceptual rationale if risk adjustment or peer grouping is not part of the measurement system. 5) Is there a conceptual rationale for the impact of social factors on scoring? 6) Was social risk adjusted considered and tested? 7) How is the risk adjustment approach consistent with the goal of the measurement system?
	Usability
	1) Are the methodologies used in the measurement system transparent? 2) Will the data be available? 3) How will performance information be displayed? 4) Who is the intended user of this information (e.g., consumers, health plans, government agency)?

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

Appendix E: Public Comment

The draft report was posted on the project webpage for public and NQF member comment on May 29, 2020 for 21 calendar days. During this commenting period, NQF received 44 total comments from 17 organizations. Comments were elicited through various avenues including the public commenting tool, an NQF town hall webinar, and additional organizational reach. NQF also received five additional comments from two individuals during the public commenting period at the close of the TEP post-comment call on July 1, 2020.

Prompts provided for public response included:

- What general comments do you have on the report?
- Does the report clearly describe the definitions and elements of measure sets and measurement systems? If not, please provide feedback.
- What future role do you feel NQF should play in advancing measure sets and measurement systems?
- Are there specific organizations that NQF should consider as potential partners or collaborators in pilot testing?
- What, if any, additional examples of sets and systems would be useful to include?

All comments received have been posted in the **comment table** on the **Measure Sets and Measurement Systems project site**. This comment table contains the commenter's name/organization, the full text comment, and its topic or theme.

COMMENT THEMES

Below is a summary of key points and high-level topics emphasized by commenters.

Definitions, Scope, and Examples

Eighteen comments from 11 commenters were received related to this topic. Several commenters agreed the report was clear on the definitions and elements of sets and systems, but acknowledged additional nuances and challenges within the field. Specific challenges mentioned include incorporating the patient and consumer voice within the discussion, the shifting of care to the primary care setting, and the need for point-of-care testing and continuity of care, more attention needed on the individual components of the measurement life cycle process, and specifications of measure sets and measurement systems and their ability to address dual and complex populations' needs. Additional themes include a suggestion to increase emphasis on both process and outcome clinical measure sets and systems in relation to patient reported outcomes. Costs were mentioned by multiple commenters, relating to value-based models, the relationship between cost and quality, and whether sets and systems include cost-related components. At least one commenter provided examples of measure sets that could be highlighted in the report.

Opportunities for Improving Sets and Systems

Seven comments were received from four organizations relating to opportunities for improving the current state of sets and systems. There were two comments focused on the measurement of patient experience in relation to sets and systems. They noted that available measures may not accurately represent the most important aspects of experience from the patient's perspective. Other comments focused on technical aspects and limitations of measurement, including measure testing and selection, methods of administration, implementation, analysis, adjustment, and reporting. There were references supporting measure sets representing a person-centered and holistic view of quality, and addressing social determinants of health and experience of care.

Sets and Systems Review Process

Nine comments were received from seven organizations related to the sets and systems review process proposed. Several comments focused on data validity, reliability, and other technical aspects of measures within a measure set or measurement system. There was support for greater transparency of these features as well as survey methods, analysis and reporting, data verification, aggregation, stratification, adjustment, and analysis. Technological advancements, including artificial intelligence, predictive models, and algorithms, were highlighted as an opportunity to generate accurate, reliable, and usable information for action. There were additional comments relating to consumer and end-user understanding and use of the quality information presented, as well as the stability of ratings for data within unstable time periods (e.g.,

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

COVID-19). Multiple commenters wanted additional clarity on the details of a multistakeholder review process and its value to the field.

NQF Collaboration and Advancement Opportunities

Seventeen comments were received from eight organizations related to NQF collaboration and advancement opportunities for measure sets and measurement systems. Several dynamic organizations offered to partner with NQF in the future of sets and systems. Several comments focused on NQF's opportunity to lead in this area by convening stakeholders throughout the healthcare industry—measure developers, government programs, private payers, patients, and consumers—to ensure all perspectives are heard when discussing measure sets and measurement systems. There were additional comments pointing toward the opportunity of technology and lifestyle medicine to be further used within the measurement landscape.

DISCUSSION OF PUBLIC COMMENTS

NQF considered all comments and discussed comments received with the TEP during a meeting on July 1, 2020. The TEP focused on the following topics:

- What is the value of an external, multistakeholder evaluation of measurement systems/quality reporting programs given the rule-making process?
- How would the review process consider programmatic, proprietary components of private or commercial programs? How can we engage creators/stewards of such sets and systems in a review process?
- How should the measure evaluation process consider "fit for purpose" (e.g., validity and reliability for specific measure score cutoffs in specific programs)?
- What are the most promising opportunities for NQF to advance this work? What are the opportunities for collaboration, specifically with public and private program stewards/developers?

The TEP affirmed that an external, multistakeholder review would be a valuable addition to the field and complementary to formal rulemaking. They highlighted the importance of an independent review and bringing together different perspectives to provide unified recommendations. Regarding measure sets and measurement systems with proprietary components used by private organizations, the TEP suggested that future outreach will be needed to determine how these organizations could best engage in a review process if not all components of the programs can be transparent. Suggestions offered included focusing on the measure set used, as the measures themselves may be more likely to be shared, and allowing reviewers only to assess the methodology, without posting it publicly.

The TEP reinforced the need for alignment of the proposed evaluation process with other NQF programs. Specifically, there is an opportunity to work with the NQF Scientific Methods Panel to determine how scientific acceptability requirements can ensure that measures are being tested for reliability and validity in the context they are intended to be used (e.g., surveillance, public reporting, payment). Additionally, if a measure or measure set is intended for broad use, members shared that reliability and validity should be tested in a population of various ages, races/ethnicities, genders, and socioeconomic characteristics. Another suggestion was to consider the development of a streamlined process to review measures when they are used for new purposes or in different populations than those originally designed and intended. There is an opportunity to explore updates to the measure endorsement process and/or address these needs through future work on measure sets and measurement systems.

The TEP expressed that NQF should advance this work by testing the framework put forth in this report against real-world sets and systems. The TEP stressed the need for a harmonized approach across systems, and noted that opportunities may emerge as the Department of Health and Human Services' National Health Quality Roadmap recommendations are enacted. Medicaid Managed Care programs were an example shared by the TEP to consider in future work. NQF looks forward to working with our members and other collaborators to further the work on measure sets and measurement systems noting these considerations.

© 2020 National Quality Forum, unless otherwise noted. All rights reserved.

NATIONAL QUALITY FORUM
1099 14TH STREET, NW, SUITE 500
WASHINGTON, DC 20005
http://www.qualityforum.org