Measurement Systems: A Framework for Next Generation Measurement of Quality in Healthcare

MARCH 2019

Mary Beth Landrum, PhD
Christina Nguyen
Erin O'Rourke
Madison Jung
Taroon Amin, PhD
Michael Chernew, PhD

Acknowledgements:
The authors thank Elisa Munthali, MPH and Shantanu Agrawal, MD, MPhil for their comments on drafts.

Supported by a grant from the Laura and John Arnold Foundation. The views presented here are those of the author and not necessarily those of the Laura and John Arnold Foundation, its directors, officers, or staff.

CONTENTS

Measurement Systems: A Framework for Next Generation Measurement of Quality in Healthcare

ABSTRACT

Performance measurement is essential for healthcare system transformation. However, the current measurement infrastructure faces important challenges that current approaches have failed to overcome. We need to understand better the scientific properties between individual performance measures and the context in which they are used. In this paper, we explore the concept of measurement systems: the interplay between a measure and its use to achieve a specific goal. The multistakeholder review of individual measures is well established. However, as measures are increasingly used to support value-based purchasing and the development of alternative payment models, stakeholders need to evaluate not just the individual measure but the overall measurement system.

INTRODUCTION

Over the past three decades, we have seen an explosion of efforts to measure the quality of healthcare. Many stakeholders, including researchers, health plans, and medical societies, have created quality measures, and the National Quality Forum (NQF) has endorsed many measures through a thorough, multistakeholder review process. Currently, NQF has a portfolio of 628 endorsed measures, but this number fluctuates as NQF both endorses more measures and removes existing measures from its portfolio.[1] These measures include structural measures (e.g., adoption of health IT tools, nurse staffing hours), process measures (e.g., HbA1c testing, amount of time between discharge ordered and actual discharge), intermediate outcome measures (e.g., HbA1c control, LDL control), outcome measures (e.g., mortality, readmission), cost/resource use measures, and patient experience measures. The measures span many clinical areas, though some areas such as diabetes care and cardiovascular care are particularly well covered. Some are calculated based on claims data; others rely on chart-abstracted data. Chart-abstracted measures require providers to pay extractors to identify data elements from a patient record and submit those data, which can be costly. Increasingly, measures also rely on electronic health record or survey data.

Quality measures are used for many purposes, including internal quality improvement and accountability applications (e.g., network design or value-based purchasing programs). Measures

are specified and applied at multiple levels of analysis such as the individual clinician, clinician group practice, facility, ACO, and the health plan level. Public and private initiatives that intend to incentivize better quality rely on these and other measures. For example, the Centers for Medicare and Medicaid Services' (CMS) Merit-based Incentive Payment System (MIPS) for clinicians is built around performance measures. Performance measures are incorporated into all CMS alternative payment models (APMs) and value-based purchasing programs such as the Medicare Advantage (MA) star rating program and the Hospital Value-Based Purchasing Program.

Parallel initiatives have expanded in commercial markets where performance measures are used for incentive payments in risk contracts. Commercial insurers also use these measures to support the creation of tiered and narrow provider networks. Quality measures are also integral to private and public efforts to facilitate beneficiary choice of health plans and providers. Finally, policymakers and analysts rely on measures to assess the impact and effectiveness of health system transformation efforts.

Challenges with the Current Measurement Infrastructure

Despite the increasing importance of quality measurement, the measurement infrastructure faces growing criticism for four primary reasons. Specifically, the current quality measurement infrastructure is incomplete, is expensive, does not ensure that measures are applied correctly, and does not calibrate performance measures for their intended use. Each of these challenges will be explored individually.

Incomplete

The current measurement infrastructure is incomplete, despite the multitudes of measures. The multidimensionality of healthcare leaves serious gaps in measurement. For example, crucial aspects of care, such as quality of diagnosis and appropriateness of care (care that yields desired results for that particular patient) are largely unmeasured and are important aspects of care in many clinical areas.

Expensive

The quality measurement system is costly in financial and nonfinancial terms. Estimates suggest that $15.4 billion is spent on collecting data for quality measures.[2] This does not count all the resources used to develop the measures. Moreover, quality measurement can distract providers from improving unmeasured aspects of quality and may get in the way of higher value care.[3,4]

Application

The current measurement infrastructure does not always ensure that measures are applied appropriately. For example, CMS is using the hospital-wide, all-cause readmissions measure which is NQF-endorsed at the facility unit of accountability in the MIPS program. This program assesses performance at the physician group level, but this measure has not been evaluated for reliability and validity at this unit of accountability. While NQF reviews thoroughly for reliability and validity in the context of the specified unit of accountability (e.g., facility) for which a performance measure is designed, it does not review for reliability or validity in specific program contexts. Measures are often used in a context outside of their endorsement.

Intended Use

Under the current quality measurement infrastructure one measure can be used for many different purposes (e.g. value-based purchasing, quality reporting) However, different uses of a measure may necessitate changes to the specification of the measure. While the current quality measure infrastructure admits the use of quality measures for many purposes, measure users often do not recognize that the developers define the specifications of the

performance measures based on their intended use. First, consumers value public reporting and information to support their decision making in selecting healthcare providers. Second, payers and purchasers are interested in using measures to affect payment in value-based contracts and to support network design. Third, providers are interested in measurement to support care improvement but are affected by all uses. For example, public reporting affects provider reputation and can impact patient volume. Value-based payment may directly impact provider reimbursement. Network design also impacts provider volume and access to patients.

Some stakeholders have argued that we should not "rank" accountability purposes, often discussing quality measurement as if there is a unique construct of quality and once specified it could be used for all purposes. Yet this is not the case. For various reasons, the nature of measurement depends on the intended use. While the current NQF process distinguishes between internal quality improvement and accountability, stakeholders are unable to fully evaluate a measure in the context of its use and provide input on several key dimensions that can impact the acceptability and value of the measure.

Statistical Issues of the Current Measurement Infrastructure

The limitations of the current measurement infrastructure have statistical implications. Since measures are used in a way that differs from the intended use for which the developers designed the measures, they may not be used in a manner consistent with their empirical testing. For example, measures may be applied at a unit of accountability different from the one for which the measure was originally specified and ultimately endorsed, and such applications impact the statistical concepts of reliability and validity. Unobserved factors, such as unmeasurable clinical differences and social and behavioral factors, can also impact measure performance, particularly when applied to small sample sizes.

Second, NQF examines the statistical thresholds of empirical testing but does not evaluate these statistical thresholds in the context of the program design (including applied unit of accountability, or sample size). The interaction between the program design and measure characteristics is often not transparent to accountable entities in the current measurement infrastructure. For example, the NQF endorsed 30-day readmission measures are used in both the Hospital Inpatient Quality Reporting Program (IRP) and the Hospital Readmissions Reductions Program (HRRP). However, these two programs use different methodologies to assess the results of the measures. While the measures ultimately calculate the rate of readmissions in both programs, they use different methodologies to interpret that result. For example, in the IQR program, CMS uses a 95 percent interval estimate to determine if a hospital's readmission rate differs from the national average. However, the HRRP uses predicted-to-expected ratios to determine a hospital's performance and penalties.[5] The current NQF endorsement and selection processes do not weigh in on such benchmarking algorithms.

Third, currently NQF often evaluates measures individually. The method in which individual measures are combined or aggregated is also relevant. For example, an important outcome measure with moderate reliability at a provider level could be combined with related process measures to improve overall quality signal, but that process of aggregation in a program is outside of the current NQF measure endorsement process.

Finally, the intended use may affect whether it is necessary to distinguish quality across the entire spectrum of performance or only identify outliers. Distinguishing among middle of the pack of providers is challenging and may not be needed for all purposes. For example, distinguishing among the middle of the pack may matter more to guide improvement rather than to guide patient choice because many providers in the middle of the pack may not be statistically different from

one another. For network construction or provider sanctioning/regulation, measuring at the tails (particularly the bottom tail) may suffice. But the statistical approach to identifying the bottom tail may differ from an approach focused on generating a score for all providers.

Moving forward, we need to rethink aspects of the quality measurement infrastructure to demand value from measurement just as we demand value from care. This thinking must go beyond the cycle of expanding measurement with new measures, followed by the search for a core measure set, and then frustration with missing measures and a call for a return to expanded measures. Specifically, we believe we need a framework for thinking about the quality measurement infrastructure. We believe that the infrastructure should acknowledge the objective or intended use of the measures and the method in which they are aggregated, and we think that the infrastructure should calibrate the statistical properties and standardization method (i.e., risk adjustment) of the measures with the incentive mechanism used in the program.

A FRAMEWORK FOR A QUALITY MEASUREMENT INFRASTRUCTURE

We believe the quality measurement infrastructure comprises three levels: measures, measure sets, and measurement systems. Measures refer to the specific aspects of performance being measured. Measure sets refer to the set of measures used for any given purpose. Measurement systems refer to how the measures are combined and used, including any other aspect of the measurement activity related to how the measures are used. In essence, measures are the list of possible ingredients. Measure sets are the shopping lists that determine which ingredients will be used in a particular application, and measurement systems are the recipes that determine how the ingredients are combined to make a meal (achieve a purpose). The recipes will vary based on purpose (taste).

Measures

A healthcare performance measure provides a way to calculate whether and how often the healthcare system does what it should. The current healthcare measurement infrastructure focuses on the development, endorsement, and use of individual performance measures to assess quality of care provided by various accountable units for individual patient populations, often defined by clinical condition.

The specifications of a healthcare performance measure generally include the following key components:

- Measure name and title
- Measure description
- Target population (denominator and numerator definitions)
- Key terms, data elements, codes, and code systems used to define the target population
- Calculation algorithm
- Timing and time intervals, if applicable
- Unit of accountability
- Data source(s)
- Sampling and stratification method, if applicable
- Risk adjustment method or exclusions, if applicable

Healthcare performance measures have been developed to suit a single unit of accountability.

For example, there are different performance measures to hold hospitals and health plans accountable for readmission rates. This is often due to the underlying data and risk adjustment approach, given the different accountable unit. The number of healthcare performance measures has increased because healthcare purchasers and providers are expanding quality improvement initiatives, and are using measures in a growing array of accountability applications. The desire for measurement to be comprehensive in terms of measuring different aspects of healthcare delivery and types of providers has also resulted in an increased number of measures assessing similar quality constructs. NQF offers endorsement for performance measures that are best in class and represent broad consensus by the healthcare stakeholder community, notably consumers. However, this process evaluates the scientific merit of individual measures without the benefit of examining these priorities within the program context in which the measures are deployed.

Criteria to Develop and Evaluate Measure

NQF endorsement uses five major criteria to assess a candidate measure for endorsement including:

1. Importance to measure: evaluates the evidence to support a measure and the potential variation in performance across providers
2. Scientific acceptability: assesses the reliability and validity of the measure
3. Feasibility: assesses the burden involved with collecting the measure information
4. Usability and use: evaluates if a measure can be appropriately used in an accountability program
5. Related and competing measures: assesses if the measure is duplicative of other measures; requests harmonization or selection of best in class

The number of NQF-endorsed measures has increased from fewer than 200 in 2005[6] to over 600 as of December 2017. This number does not include measures that are not endorsed, either because they failed an endorsement review or have never been submitted for review but are still in use in federal or private payment programs.

Issues at an Individual Measure Level

Performance measures must be accurate and meaningful if they are to drive behavior change and performance improvement. The foundation of an accurate and meaningful measure is the evidence to support it. Before assessing a measure's statistical merits, the NQF review process assesses if a measure is important to measure and report. This criterion is meant to assess if the measure is evidence-based and important for improving healthcare quality. However, certain challenges have emerged in assessing the importance of a measure.

First, the connection between improving processes and improving outcomes is not always clear, making the link between process measures and patient outcomes less robust than optimal. As part of its review process, NQF examines the evidence to support endorsed measures. Specifically, for process measures, NQF requires a review of the quality, quantity, and consistency of the published evidence that the process intervention affects the outcome. However, if the evidence is unavailable, NQF allows for a systematic assessment of expert opinion, or clinical guidelines, that indicates that the benefits of measurement outweigh any harms. Moreover, NQF requires empirical evidence for outcome measures linking the outcome to at least one healthcare action. However, despite evidence to support both the processes and the outcomes, the correlation between improvement on process measures and improved outcomes has been a central debate in quality measurement.[7] There are a myriad of interventions a provider could undertake to influence a patient's outcomes, but many may not influence an individual process measure used to assess quality. Moreover, many interventions to improve this process measure

may not have corresponding impacts on patient outcomes. Multiple process measures may needed to capture the multidimensionality of the patient outcome of interest.

Another potential challenge to the acceptance of measures is the opportunity for meaningful improvement. Individual measures that are "topped out" may not be meaningful. A topped out measure is one with high levels of performance with little variation and, therefore, little room for further improvement. For example, almost all providers may demonstrate very high performance on process measures such as hemoglobin A1c exam or eye exam. With lack of variation, provider performance would fall into a tight distribution around the mean, making it difficult to distinguish between providers of low or high quality on that measure. Similarly, crossing a certain threshold or moving several percentiles in ranking in a tight distribution may misrepresent the magnitude of difference between levels of performance. The lack of variation potentially reduces the meaningfulness of these measures. While the NQF process reviews measure performance to determine if a measure is topped out, this process relies on provider performance in a test population determined by the measure developer. A measure may be topped out in a specific application with a smaller set of providers. How an individual topped out measure is used with other performance measures to capture an outcome of interest and the method of aggregation of these multiple measures may determine if it is important to continue to use a topped out measure. For example, the topped out measure may be used as a monitoring tool for unintended consequences to patients.

We also need to consider the attribution of the measure result. The attribution model at the measure level identifies the individual patients who will be included in the denominator of the measure, the accountable unit of the measure result, and the data used to determine the provider and patient relationship. An examination of the attribution model for an individual measure should consider the degree of control a provider has over a healthcare outcome, if the sample size is sufficient to ensure reliability, and if the risk adjustment model and measure exclusions allow for comparable patient populations.

Statistical issues exist at the individual measure level. First, measures are susceptible to reliability and validity errors. Reliability is the repeatability of measurement. Reliability is a largely driven by sampling variation and thus small sample sizes, high within-provider variation, or low across-provider variation can each lead to low reliability. The issue of small numbers has impacted the reliability of performance measures assessing rare events.

Validity is the correctness of measurement as compared to an authoritative source. The validity of a performance measure could be tested by testing hypotheses that the scores indicate quality of care, (e.g., scores are higher for groups known to have better quality assessed by another valid quality measure or method); the correlation of measure scores with another valid indicator of quality for the specific topic; or a relationship to conceptually related measures (e.g., scores on process measures correlate to scores on outcome measures). Errors of validity may be due to variation in the infrastructure and methods used for measurement based on setting and provider (e.g., coding practices, structures and processes in place for documentation). Reliability and validity go hand-in-hand: when measurements are unreliable, performance can be incorrectly categorized, which results in loss of validity. The NQF endorsement process evaluates candidate measures for reliability and validity; however, this evaluation is often not in the context of the specific program population or calibrated to the program intent. NQF does not dictate neither the statistical test nor minimum thresholds for results of reliability or validity testing since measure developers have a multitude of testing situations. For example, if a measure developer is empirically testing the reliability of a measure score using a spilt sample reliability test, NQF does not prescribe

the expected kappa scores. This lack of thresholds makes the calibration of the statistical properties of healthcare performance measures to specific levels of accountability, or financial risk in value-based arrangements difficult, if not impossible. In the endorsement of measures, NQF evaluates the statistical risk adjustment model and risk model performance as part of the specification of outcome measures (clinical or economic outcomes). This type of adjustment is critical to ensuring an appropriate comparison of provider performance. While adjustment for patient clinical risk factors has been the standard for outcome measures, the appropriateness of adjustment for social risk has been the source of much debate. Prior to 2015, NQF prohibited the inclusion of social risk factors due to concerns about masking healthcare disparities. However, the increased desire to use outcome measures in accountability purposes led NQF to explore this issue. NQF ultimately concluded that social risk factors can be included in a risk adjustment model on a measure-by-measure basis when there is a conceptual basis and empirical evidence to support their inclusion. Recent work has demonstrated that the statistical model used for risk adjustment is important.[8] Specifically, is it important to distinguish between within-provider associations between social risk factors and quality and sorting of patients with high social risk factors to low quality providers.

Measure Sets

In order to evaluate provider performance on multiple measures, some form of aggregation is needed. Measure sets are the first step in aggregation. Groups of individual measures form sets, often created based on intent. A measure set could refer to a group of measures intended to work together or a pick-list of measures from which to select.

It is important to distinguish measure sets from composite measures. NQF defines a composite measure as a combination of two or more individual measures into a single measure that results in a single score. Composites are constructed through five steps: (1) identify the purpose and the quality construct to be measured, (2) select the measures and/or subcomposite measures to be combined, (3) ensure that the weighting and scoring of the components supports the goal that is articulated for the measure, (4) Combine the component scores, using a specified method, into one composite (e.g., sum, average, weighted average, patient-level all-or-none scoring), and finally, (5) test the measure to determine reliability and validity.

Measure sets do not typically create a composite score for an accountable entity from the component measures used. The creation of a measure set focuses on the first steps of identifying a purpose, defining a quality construct, and selecting measures to assess that quality construct but does not involve combining the measures into one score or testing for reliability or validity.

Broadly, measure sets can refer to compilation of individual measures for the following three uses:

1. Defining high quality clinical care in a disease area; or,
2. Defining high quality care for an accountable unit or setting; or,
3. Defining how to advance health system priorities, such as safety, patient engagement across settings or clinical areas

Importantly, these three uses may not be mutually exclusive.

Measure sets have been used to define high-quality clinical care. For example, the ORYX measures developed by The Joint Commission are chart-abstracted measures in clinical topic areas designed to support quality improvement initiatives in that area. This approach is clinically appealing as it can capture the multiple interventions or elements of quality care, defined by evidence that leads to improved patient outcomes.

Measure sets have also been used to define high-quality care for an accountable unit or setting. For example, the Consumer Assessment of Healthcare Providers and Systems (CAHPS) is a series of surveys in which patients rate different aspects of their care.[9] Further, the Centers for Medicare and Medicaid Services (CMS) created a measure set in 2014 to evaluate the quality of care provided by ACOs to Medicare patients as part of the Medicare Shared Savings Program. The 33 measures were further separated into the four domains patient/caregiver experience, care coordination/patient safety, preventive health, and at-risk population. All of these domains are intended to define good care delivered by an ACO.

Finally, measure sets can also be developed to advance health system priorities (safety, patient engagement) across settings or clinical areas. For example, in the recent public health epidemic, measure sets for opioids-related care can be developed to incentivize multiple care sites and clinical specialties to focus attention on the epidemic. The development of a measure set in this area can help identify new roles that various portions of the healthcare delivery system can play to impact health system priorities.

Criteria to Develop and Evaluate Measure Sets

Unlike individual performance measures, there is no standard process to develop a measure set. Additionally, there are no standard criteria to define or evaluate a measure set. A single measure developer could create a measure set for the purpose of assessing multiple domains of quality in one aspect of healthcare, for example the CAHPS and ORYX measure sets noted above. Alternatively, a group of stakeholders could select a measure set containing measures from multiple developers. The National Academies of Medicine (NAM) issued the Vital Signs report recommending a core set of concepts that should be used to develop a measure set. Additionally, NQF convenes the Measure Applications Partnership (MAP) workgroups to review measure sets. These groups have developed their own processes and criteria for selecting measures for a measure set.

As measure sets continue to gain popularity as an approach to promoting alignment and reducing measurement burden, there may be a need for review of measure sets against a set of standard criteria. Evaluation of measure sets could promote standardization and ensure the measures in the set adequately address quality for the construct being measured. Measure sets are the first step in aggregating measures to make inferences about provider quality, so there is a need for increased transparency through a multistakeholder review on how the sets are developed, how quality is defined, and how the sets are designed to work.

Issues for Measure Sets

Creating measure sets, or lists of high-priority measures, has gained popularity as a way to reduce the burden of measurement and make sense out of which of the hundreds of available performance measures to use.

While a valuable tool for burden reduction, the focus on measure set development has important limitations. First, activities to develop or evaluate measure sets do not always look at how measures in the set work together. A notable limitation to measure sets is that there may not be a requirement to use all measures in the set together. Some sets are designed as a “pick list” rather than a comprehensive set. For example, to meet the quality domain, MIPS allows clinicians to select six measures to report. While this approach may ensure that clinicians feel that measures reflect their practice, the “pick list” approach to a measure set creates scores that may not truly represent a complete patient episode of care and raises concerns about the ability of the set to facilitate comparability of measurement between providers.

Additionally, measure sets are only an effective tool to reduce the burden of measurement if

there is broad stakeholder agreement on the measures in the set and how and where they are implemented. Stakeholders may disagree on whether to use all of the measures in a set. Reducing measures in one set only reduces the administrative burden of measurement if the excluded measures are not included in other sets. Moreover, measure implementers may not implement the measure exactly as specified, introducing variation and negating the goal of administrative burden reduction.

Finally, efforts to develop measure sets do not seek to evaluate how the individual measures roll up into a composite score, or how the individual measures are weighted in an accountability program's scoring algorithm. For example, MAP recommends individual measures to add to the Hospital Value-Based Purchasing Program. This effort is important in gathering multistakeholder feedback on the selection of an individual measure to include in the program measure set. However, this program groups measures into domains to determine a hospital's final score. MAP does not comment on what domain a measure should be in, what other measures should be in that domain, or how the domains are weighted in the final scoring algorithm.

Measurement Systems

Measurement systems refer to how measures are used to achieve a goal. Measurement systems vary by context, setting, and intended use. Despite the variation, key elements define a measurement system. First, there is the objective of the measurement system: what cost or quality issue is the system trying to improve? The method of aggregation is a critical element of a measurement system that includes methods for standardizing scales across component scores, weighting rules, handling of missing data, and required sample sizes. Next, there is the incentive mechanism the system will use to drive improvement (e.g., public reporting, value-based payment, or capitated payment). Finally, a measurement system can include a risk adjustment approach to standardize the population being measured in the system. Measurement systems combine these aspects to make inferences about performance of a provider or a policy.

Objective of the Measurement System

The objective of the measurement system and the intended use of the measures are the first elements. For example, one possible use is to encourage providers to improve the quality of their care. To accomplish that, detailed information about the activities that lead to higher or lower quality can be helpful. Thus measurement for quality improvement (QI) purposes would seek specific measures and potentially focus on process measures or proximate outcome measures. Reporting high all-cause readmission rates may be too general to promote quality improvement, as providers need details about which conditions are driving the high rates, so interventions can be tailored effectively. While condition-specific measures can be challenged by sample size and confounding, statistical noise may be more tolerated in QI-only programs, as the main goal of the measures is to identify potential areas that need improvement and direct improvement activities accordingly. These programs are often internal, and results are not shared publicly, so the risk of misclassification is not perceived to be as great as when results are used to inform provider selection or determine payments.

A quality measurement system may also support patient choice of providers using public reporting of quality performance. This also requires measurement of specific providers and procedures, as well as adequate case mix adjustment. However, supporting patient choice requires a focus on outcomes meaningful to patients that, in many cases, may be hard to observe. Moreover, we may be less accepting of noise when we are informing patients through public reporting, as statistical errors can impact a provider's reputation and a patient's choice. However, measures of important patient outcomes with low to moderate reliability individually

could potentially be combined with other related measures leading to improved statistical properties and clinical relevance.[10,11]

Measurement systems may also determine provider payments. In theory, measurement for the purposes of rewarding a health plan, delivery system, or ACO is in many ways easier because aggregation across providers in the system and across conditions addresses the sample size issues. This would include pay-for-performance programs or use of quality measures for network construction, provider tiering, or reference pricing. In some of these cases, the degree of reward (or penalty) can be titrated to the precision of measurement or the clinical importance of the measure. However, this is not currently done in practice. For example, process measures are used in CMS' Hospital Value-Based Purchasing (VBP) Program (with a proposed weight of 50 percent of the total score), while programs like its Readmissions Reduction Program or Hospital-Acquired Condition Reduction Program (HACRP) use only outcome measures that some stakeholders would argue do not have adequate reliability scores to distinguish performance among providers.

Finally, a measurement system can draw broad inferences about programs. Do ACOs improve quality? This use may be the easiest because the sample size is greatest. The main takeaway is that a quality measurement system should be calibrated to the objective, especially when measurement is at a more granular level (physician, versus group versus delivery system/ ACO versus health plan) and the intended use entails greater sanction. The willingness to accept imprecision in measurement diminishes as the sanctions rise.

Method of Aggregation

Existing quality measurement systems focus on aggregation of measures into scores, either overall or by domain and then overall. Issues arise in this process including whether the grouping and weights reflect statistical properties or normative values.

Measurement systems often use an ad hoc approach to aggregation, developing complicated weights and setting payment functions. For example, MIPS combines requirements from the Physician Quality Reporting System (PQRS), the Medicare EHR Incentive Program (Meaningful Use), and the Value-Based Payment Modifier. Providers are evaluated on the performance categories, or domains, quality, advancing care information, improvement activities, and cost. MIPS uses a unified scoring system that converts performance on the individual measures in each performance category into points. Each performance category gets a weighted value in computing the composite performance score (ranging from 0-100 points). MIPS makes negative, neutral, and positive payment adjustments on the composite performance score based on failure to meet performance thresholds for exceptional performance.

Many quality measurement systems such as CMS' ACO Program use normative grouping (e.g., based on clinical relationships) and weighting (e.g., equal weighting within domains and across domains to produce overall composite scores). While common, such normative approaches could mask important aspects of underlying quality by combining a provider's scores for two different dimensions of quality (e.g., diabetes control and depression screening). An alternative, empirical approach could use statistical relationships between measures to create groupings and weightings used to compute an overall composite score.[12]

Currently, the details of how the aggregation approach was developed may not be available to all stakeholders, which points to a need for greater transparency. The CMS Hospital Star Ratings offer one recent example of the need for greater understanding of how measures are aggregated. The Star Ratings methodology summarizes data from performance measures reported on the Hospital Compare website into an overall rating to simplify the information for consumers. However, concerns have arisen about how these measures

are aggregated into the overall rating. For example, measures are assigned to a domain and the domains are assigned different weights within the overall score. The variation in the weighting of domains can influence overall performance. Hospitals that perform well on heavily weighted domains are more likely to achieve a five-star rating.[13] However, the decision making process behind the domain weights affects the outcome and is not transparent to all stakeholders. Moreover, factors outside a hospital's control can influence ratings, including the ability to report a measure and the influence of the underlying case mix of the patients the hospital serves. Finally, changes in the underlying methodology can result in changes in performance not driven by changes in quality. For example, through the updated methodology used in 2018, three times as many hospitals received a five star rating as in 2017.[14]

The influence of these factors emphasize the need for a multistakeholder review. Key elements of aggregation that should be considered include:

- Component measures are well defined and precisely specified;
- Defined methods for standardizing scales across component scores
- Scoring rules (i.e., how the component scores are combined or aggregated)
- Weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite)
- Handling of missing data
- Required sample sizes.
- Statistical properties of aggregate scores

Incentive Mechanism

In systems focused on rewarding performance, the measurement system often translates the aggregated score into a financial reward. This function that relates the score to rewards is typically not linear. For example, MIPS uses performance rankings to determine how to redistribute the pool of bonuses. Thus, for every winner, there is a loser. In some cases, rewards or penalties only apply to those organizations that perform in the upper tail (as in the exceptional performance portion of MIPS, or those in the bottom tail as in the Medicare ACOs that do not meet saving targets). Moreover, in some cases, those in the bottom tail may incur some other sanction that may not be directly financial. For example, the Hospital Inpatient Reporting Program collects data, some of which is reported on the Hospital Compare website. Hospitals that score in the bottom percentages on a measure have their performance reported as lower than the national average on that measure, which could negatively impact their reputation and encourage consumers to choose another provider.

Provider performance rankings can be particularly sensitive to thresholds. Future measurement systems can consider thresholds to determine whether providers perform well enough to be rewarded (thresholds for minimum attainment are often set at the individual measure level). In order to create these thresholds, providers should be ranked based on their performance. Then, depending on the context, a threshold is determined. In programs such as MIPS where rewards are based on ranking, a slight change in threshold can move a provider up or down (and when one goes up, another must go down).

The development of measurement systems needs to calibrate measures to the intended use or accountability application. For example, in the Hospital Readmissions Reduction Program (HRRP), a multistakeholder group should review the statistical calibration, or the degree of statistical confidence of the individual measures in relation to the nature of the penalty.

In systems focused on reporting, decisions must be made about how the results are reported to consumers. For example, systems could report absolute scores or could report category ratings derived from the scores. A key challenge to

reporting to consumers is ensuring consumers can understand the information. Performance measures and in turn, measurement systems, produce complex statistics that a layperson may find difficult to interpret. However, attempts to use plain language or to use simpler reporting mechanisms like stars rather than numbers have caused controversies. Stakeholders have expressed concerns that these results do not truly indicate a provider's performance. Other approaches, such as only reporting outliers, have drawn criticism for not providing enough granularity to support consumer decision making.

Risk Adjustment

Another statistical challenge is that healthcare outcomes do not result solely from healthcare interventions but involve time and patient factors as well. Often, meeting the performance target requires both physician and patient action, but patients are influenced by external factors such as insurance benefit design and their desire for care when and where they want it. While risk adjustment attempts to address the effect of patient factors, there is limited consensus on the use of social risk factors, rather than only clinical factors, in current models. Further, there is limited robust data available on social risk factors that influence the outcomes being measured.

Risk adjustment, particularly social risk adjustment, can be made at the individual measure or aggregate level (i.e., domain/set, overall quality score). For example, VBP adjusts individual measures, while MIPS adjusts based on providers' composite performance scores (MIPS has set maximum adjustments).

Directly adjusting the measure allows consistent use of a single measure across multiple use cases. For example, stakeholders have questioned if the CMS 30-day readmission measures used in both the Hospital Inpatient Quality Reporting Program and the Hospital Readmissions Reduction Program should include social risk factors in their risk adjustment models. Including the factors in the risk adjustment model of the measure may allow for consistent and comparable measure results across programs.

Other approaches to comparative performance assessment, such as stratification or peer group comparison, happen at the measurement system level. Stratification refers to computing performance scores separately for different groupings or strata of patients based on selected characteristics. Essentially, each healthcare unit receives multiple performance scores (one for each stratum) rather than one overall performance score. Peer group comparison involves creating peer groups for providers caring for a similar mix of patients and examining scores within that group. However, it is important to note that statistical risk adjustment, stratification, and peer groups for comparison are not mutually exclusive. These approaches could be used in various combinations or in all three ways for a given performance measure, with the specific analytic approach chosen for a specific analytic or program purpose.[15]

Attribution Model

Attribution is a methodology to assign patients, encounters, or episodes of care to a healthcare provider or practitioner. An attribution methodology seeks to determine the relationship between a patient and his or her team to ensure that the correct entity or entities are accountable for the patient's outcomes and cost. As noted above, attribution can be an element of the design of an individual performance measure as well as a measurement system.

As part of a measurement system, attribution outlines the rules for assigning patients to the accountability program. The attribution model for a measurement system should align with the goals of the system, as attribution is a powerful tool to drive accountability for outcomes. A measurement system should evaluate the specific rules and methods of the various measures used for attribution. There are varying attribution methods currently performed, and there is a lack of

objective evidence to recommend one approach over another, necessitating multistakeholder review of the model as part of the overall evaluation of the measurement system.

Future Considerations: Adaptive Systems

While existing measurement systems tend to follow a relatively straightforward path from data collection to aggregation to scoring or reporting, more advanced measurement systems should be explored to help reduce the large administrative and financial burden that quality measurement places on providers. One such approach is an adaptive or targeted approach. Rather than requiring all providers to collect costly data all measures, an adaptive approach would target certain accountability units for further data collection based on easy-to-gather measures, such as administrative and claims-based measures, or electronic medical record (EMR) data. In such a model, the data collected from each organization may differ depending on its performance on common, easier to collect measures. For example, adaptive systems may aim to identify low-performing providers from which to audit or require more data and/or adjust performance with socioeconomic factors. Similarly, high-performing providers could be given more lenience in the amount or types of data they are required to provide; if shown to be consistent, data from these providers could also be collected less frequently. High performers can serve as models for effective practices.

Moving towards a system approach, such as an adaptive system, may solve many of the current challenges in measurement infrastructure. In an adaptive approach, many accountable units may be exempted from further scrutiny providing relief from measurement burden. Collecting detailed information on a subset of accountable units aims to address incompleteness of core measurement sets. Which and how many units are targeted for further data collection depends on intended use of the system.

In addition, adaptive systems can be designed to account for statistical properties of quality measurement. For example, many easy-to-gather measures may have small sample sizes, be based on relatively rare outcomes, or be subject to unmeasured confounding. These issues can be particularly problematic when classifying providers, as statistical noise or unmeasured confounders may inaccurately identify some providers as being low- or high-quality. While mislabeling some providers as low- or high-performing may impose a cost, this cost is likely lower than that of missing poor care or rewarding providers who are not truly high-quality. An adaptive measurement system may not accurately classify providers 100 percent of the time, but can be modified depending on the intended use of the system and required level of stringency to minimize costs of misclassification errors.

Adaptive measurement systems may offer a promising solution to many measurement challenges, but this approach would benefit from a multistakeholder review to consider potential drawbacks and unintended consequences. For example, patients may wish to see more granular performance data to inform their decisions about which healthcare provider to choose. However, methodologies that only identify outliers will not provide such data. Additionally, clinicians and providers may see value in data that is granular enough to identify root causes of a quality problem and allow for targeted improvement activities. Receiving information that their performance does not vary from the average may not help them to achieve their quality improvement goals.

PATH FORWARD

Currently, the measurement infrastructure enterprise focuses on the merits of individual measures. However, greater consideration and study of measurement systems could help to address current challenges to performance measurement. The authors propose the following set of recommendations as a path forward.

RECOMMENDATIONS

1. The quality measurement framework should consider measure sets and measurement systems explicitly.
2. The current proliferation of measure sets and their role in defining quality and serving as the first step in aggregation necessitates an approach to assess them systematically.
3. The quality measurement community must devote more attention to the development of the science of measurement systems.
4. Best practices for measurement systems must be defined, and criteria must be developed to evaluate measurement systems.
5. A mechanism, such as a multistakeholder evaluation process, must be established to ensure transparency and scientific soundness of measurement systems.

REFERENCES

1 National Quality Forum. *NQF Report on 2017 Activities to Congress and the Secretary of the Department of Health and Human Services*. Washington, DC: National Quality Forum; 2018:102. https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=87189.

2 Casalino LP, Gans D, Weber R, et al. US Physician Practices Spend More Than $15.4 Billion Annually To Report Quality Measures. *Health Aff (Millwood)*. 2016;35(3):401-406.

3 Berenson RA, Rice T. Beyond Measurement and Reward: Methods of Motivating Quality Improvement and Accountability. *Health Serv Res*. 2015;50:2155–2186.

4 Mullen KJ, Frank RG, Rosenthal MB. Can you get what you pay for? and the quality of healthcare providers. 2014;41(1):64–91.

5 Centers for Medicare and Medicaid. *Readmissions Reduction Program (HRRP)*; 2018.

6 Meyer GS, Nelson EC, Pryor DB, et al. More quality measures versus measuring what matters: a call for balance and parsimony: Table 1. *BMJ Qual Saf*. 2012;21(11):964-968.

7 Werner RM, Bradlow ET. Relationship between Medicare's Hospital Compare performance measures and mortality rates. *JAMA*. 2006;296(22):2694–2702.

8 Roberts ET, Zaslavsky AM, Barnett ML, et al. Assessment of the Effect of Adjustment for Patient Characteristics on Hospital Readmission Rates: Implications for Pay for Performance. *JAMA Intern Med*. 2018;178(11):1498.

9 About CAHPS | Agency for Healthcare Research & Quality. https://www.ahrq.gov/cahps/about-cahps/index.html. Last accessed March 2019.

10 Dimick JB, Staiger DO, Hall BL, et al. Composite Measures for Profiling Hospitals on Surgical Morbidity: *Ann Surg*. 2013;257(1):67-72.

11 Dimick JB, Staiger DO, Osborne NH, et al. Composite Measures for Rating Hospital Quality with Major Surgery. *Health Serv Res*. 2012;47(5):1861-1879.

12 McDowell A, Nguyen CA, Chernew ME, et al. Comparison of Approaches for Aggregating Quality Measures in Population-based Payment Models. *Health Serv Res*. 2018.

13 Understanding Medicare's 5-star rating hospital program - Modern Healthcare. https://www.modernhealthcare.com/article/20170828/SPONSORED/170829897. Last accessed February 2019.

14 Morse S. *CMS Posts Star Ratings and Three Times as Many Hospitals Earned 5 Stars*; 2018. https://www.healthcarefinancenews.com/news/cms-posts-star-ratings-and-four-times-many-hospitals-earned-5-stars-0.

15 National Quality Forum. *Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors*; 2014.

978-1-68248-110-3