American Urological Association

Healthcare Performance Measurement: A Primer for Urologists

Karen Johnson, PhD
AUA Director of Quality & Measurement

Gregory Auffenberg, MD, MS
AUA QIPS Committee Chair

David Friedlander, MD, MPH
AUA QIPS Committee Member

May 2024

Contents

Introduction

Healthcare performance measures are tools used by various stakeholders to quantify aspects of care delivered to patients and their families, including those related to quality, access, resource use, and cost.[1] As such, performance measurement in medicine helps us understand the impact of healthcare delivery on the lives of patients and use this understanding to identify ways to improve care and outcomes.

Healthcare performance measures typically quantify aspects of care for specific entities (i.e., the "accountable entity"). Usually this is done by aggregating patient-level data to reflect the performance of that entity. The accountable entity may be a direct provider of care (e.g., an individual clinician or clinician group) or a collective responsible for the provision of care (e.g., a clinic, hospital, nursing facility, home health agency, hospice agency, pharmacy, etc.).

Measurement is an improvement tool, not an end in and of itself

The primary goal of healthcare performance measurement is to improve care for patients and their families, and ultimately, to improve health outcomes.

In contrast, population-level performance measures provide information about specific aspects of care for a particular group of individuals (e.g., total U.S. population or other geopolitical areas such as a state or community). As such, these measures are not designed to hold a specific entity accountable for care delivery. Population-level performance measures may focus on various aspects of the healthcare delivery system such as access or utilization or on other important topics (e.g., health behaviors; community-based services).

This Primer provides an overview of fundamental concepts in healthcare performance measurement, with examples that are specific to urology. In addition, the Primer briefly introduces several advanced topics that often arise during measure development and implementation.

Measurement Basics

This section describes key "basics" of healthcare performance measurement, including how measures are used, categories and types of measures, data used in measures, and the measure lifecycle.

Uses of Healthcare Performance Measures

Although there are many ways in which healthcare performance measures are used, the primary goal of measurement is to ***improve*** the care that is provided to patients, and ultimately, to improve health outcomes. Healthcare performance measures also are used in a variety of accountability applications.

Measures are used for improvement activities by providing information about "where you are" in some aspect of care. If measure results indicate there is room for improvement, stakeholders may decide to initiate an improvement activity. This will include some type of intervention, or set of interventions,

[1] Primary source materials for this document include National Quality Forum evaluation criteria and guidance materials (https://www.qualityforum.org/show_content.aspx?id=322) and CMS Measure Management System blueprint resources (https://mmshub.cms.gov/about-quality/MMS-overview).

often initially implemented in a small way (e.g., for one team, one shift, one floor, etc.). After some period of time, the relevant measure(s) are re-calculated, and those results will indicate whether the changes generated improvement in the aspect(s) of care that were targeted (note that a change does not necessarily result in improvement).

In accountability applications, an accountable entity's measure results are compared against some standard, usually with implications for the accountable entity based on that comparison. Major types of accountability applications include accreditation or certification programs, contracting arrangements, public reporting, and payment programs. Note that the standards, or comparators, used in accountability comparisons will vary by program. Common comparators include the average measure result across some group of peers or across some group of recognized leaders.

Accrediting and certification bodies may use measure results to determine whether a healthcare entity meets a particular level of quality as a prerequisite for certification or accreditation. Well-known examples include accreditation of hospitals by The Joint Commission and accreditation of health plans by the National Committee for Quality Assurance (NCQA). Similarly, those who pay for healthcare (e.g., employers, government programs, health plans) may use measures to inform various contracting decisions such as inclusion in an accountable care organization or health plan network.

Various entities may make accountable entities' measure results available to patients and other stakeholders. The goal of such public reporting programs is to inform those who consume healthcare services, thereby facilitating choice. For example, a patient may compare the performance of several urologists prior to deciding which practice, or which clinician in a practice, to consult. Care Compare, a well-known resource maintained by the Centers for Medicare & Medicaid Services (CMS), reports the results of many different measures across a variety of healthcare settings and accountable entities.

Although accrediting, contracting, and public reporting can affect clinician remuneration indirectly, payment for care may be explicitly impacted, positively or negatively, based on performance measure results. For example, in the Merit-Based Incentive Payment (MIPS) pathway of the CMS Quality Payment Program, urologists' fee-for-service Medicare Part B payments are subject to bonuses or penalties based on scores determined, in part, on their performance (relative to their peers) on various quality and cost measures. CMS implements additional pay-for-performance programs that apply to other healthcare settings (e.g., hospitals) or models of care (e.g., accountable care organizations). Similar payment incentive programs also exist among private payers.

Categories of Healthcare Performance Measures

Healthcare performance measures can be classified into four broad categories based on the overall aspect of care they strive to assess, as follows:

- Quality of care
- Access to care
- Resource use/cost
- Efficiency

However, there is no clear consensus on the definitions of these measure categories. Generally, quality measures assess care safety, timeliness, effectiveness, efficiency, equity, and patient-centeredness (i.e., the dimensions of quality identified by the Institute of Medicine in its 2001 *Crossing the Quality Chasm*

publication). Access measures typically focus on issues related to availability, accessibility, affordability, and convenience of healthcare resources. Resource use measures count the frequency of use of services. Cost measures go beyond simply counting services by applying a dollar amount to each resource unit used. Finally, efficiency measures are yet-to-be-developed measures that "marry" cost or resource use measures with quality measures to reflect cost or resource use associated with a specified level of health outcome.[2]

Stakeholders do not always agree on what category a particular measure falls into because the aspects of care the categories represent are not mutually exclusive. For example, resource use often is viewed as a "proxy" for quality (e.g., too many readmissions may reflect poor quality care). Nonetheless, it is helpful to differentiate the various categories of measures, both to highlight conceptual differences and for evaluation purposes. For example, many stakeholders believe that one cannot have high quality healthcare if there is not access to care; however, one might have access to care that is not of high quality. Similarly, cost is qualitatively different than quality, and lower-cost care does not necessarily mean lower-quality care (or vice-versa). Moreover, there may be reasons to evaluate a measure that assesses access or cost differently than one might evaluate a measure that assesses care quality. For example, issues of attribution are of particular concern for cost measures, while the need (or not) for risk adjustment is particularly relevant for access measures.

Types of Healthcare Performance Measures

As with categories of measures, there are a variety of measure types, although there is not always agreement among stakeholders regarding a particular measure's type. Moreover, as the measurement enterprise evolves, consensus around a schema for measure types also may change. Nonetheless, most agree that healthcare performance measures generally can be grouped into three main types: **structure, process, and outcome**. These were articulated initially by Avedis Donabedian in 1966[3] as "approaches to the acquisition of information about the presence or absence of the attributes that constitute or define quality." Many stakeholders further differentiate various "subtypes" of outcome measures, including those that assess health outcomes, intermediate clinical outcomes, and other aspects of health or healthcare using information reported by the care recipient.

As indicated by the definitions provided below, structure, process, and outcome measures are most common in quality measures. Conceptually, however, these measure types also could be located within the access category and potentially within the resource use/cost category. For example, measures may assess structures (e.g., extended evening hours) or processes (e.g., help making follow-up or referral visits) that lead to improved access to care.

STRUCTURE MEASURES

Structure measures assess the conditions under which care is provided (i.e., the "infrastructure" of care) or characteristics of those providing care. Stated differently, structure measures assess aspects of the healthcare system relevant to its capacity to deliver good care. Such measures can provide

[2] This definition is similar to what some think of as "value", although the term "value" typically accounts for stakeholder priorities and preferences as well.

[3] Donabedian, A. The definition of quality and approaches to its assessment. In: Explorations in quality assessment and monitoring, Volume 1. Ann Arbor MI: Health Administration Press. 1980 (p.90).

valuable information about institutional capacity or capabilities, staffing, and the volume of procedures performed by a provider.

Examples of concepts that could be assessed via structure measures include:

- Level of subspecialty training
- Adoption of medication e-prescribing
- Procedure volume (hospital and/or surgeon-level)
- Participation in a clinical data registry such as AUA's AQUA Registry

PROCESS MEASURES

Process measures assess whether a healthcare-related activity (or set of activities) has been performed for, on behalf of, or by a patient. Process measures often focus on things that should be done because there is evidence that doing so will lead to desired outcomes. However, some process measures focus on things that should not be done, because there is evidence that doing so will lead to undesired outcomes.

Examples of concepts that could be assessed via process measures include:

- Conduct of a patient-centered surgical risk assessment
- Urinalysis conducted prior to invasive urologic procedure
- Timely repeat TURBT in stage T1 bladder cancer patients
- Inappropriate imaging (e.g., wrong type, frequency, timeframe)
- Timely receipt of BCG after bladder cancer diagnosis

OUTCOME MEASURES

Outcome measures assess consequences resulting from the delivery of healthcare. There are three main "subtypes" of outcomes that can be measured: health outcomes, intermediate clinical outcomes, and patient-reported outcomes.

Health outcome measures assess a patient's health status or change in health status due to some healthcare intervention or set of interventions. These are usually assumed to be measured "objectively", although certainly health outcomes can be reported by individuals (thus illustrating an overlap in this measure type schema). Unlike many process measures, outcome measures often focus on things individuals particularly care about, such as death, complications, or functional ability. For example, a patient isn't likely to be too concerned about whether he had a urinalysis prior to a surgical stone procedure, but he would be concerned if he contracted a post-surgical infection.

Examples of concepts that could be assessed via health outcome measures include:

- Injury during a hospital stay
- Infections following prostate cancer biopsy
- Hospital readmissions (e.g., if assumed to be a proxy for poor discharge planning)

Intermediate clinical outcome measures assess circumstances or changes that lead to longer-term health outcomes due to healthcare intervention(s); typically, although not always, these will be changes in physiologic state. Such outcomes may be of greater interest to patients than many care processes, but likely are not as important to patients as those focusing on health outcomes. For

example, patients may be somewhat concerned with their blood pressure readings—but more likely, they are most interested in the risk of, or occurrence of, a stroke or heart attack.

Examples of concepts that could be assessed via intermediate clinical outcome measures include:

- Blood glucose control
- Blood pressure control
- Excessive radiation dose

Outcome measures based on patient recount (i.e., ***patient-reported outcome based performance measures, or PRO-PMs***) assess some aspect of a person's health based on information coming directly from that person, without interpretation of the response by a clinician or anyone else. Such measures typically focus on topics such as health-related quality of life, functional status, symptoms, and symptom burden. Note that the term "patient-reported" is a term of art, and the term "person-reported" would be more appropriate given that patients are not the only consumers of healthcare. Note also that PRO-PMs can extend beyond clinical care (e.g., supportive services provided to those with disabilities).

Examples of concepts that could be assessed via PRO-PMs include:

- Change in depression symptoms
- Change in International Prostate Symptom Score
- Pain intensity
- Sexual function following prostate cancer treatment
- Goal attainment post-surgery

Note that it is important to differentiate the following terms related to PRO-PMs:

- The patient-reported outcome (PRO) is the topic of interest (e.g., depression symptoms)
- The patient-reported outcome measure (PROM) is the instrument used to collect data about the outcome of interest (e.g., the PHQ-9 tool)
- The patient-reported outcome based performance measure (PRO-PM) aggregates patient data collected from a PROM for some accountable entity such as a clinician practice

- Example PRO-PM: Percentage of patients with diagnosis of major depression or dysthymia and initial PHQ-9 score >9 with a follow-up PHQ-9 score <5 at 6 months

OTHER TYPES OF MEASURES

While not technically PRO-PMs, other ***instrument-based performance measures*** rely on data reported by individuals, typically in response to a questionnaire or survey. Perhaps the most common are those assessing *patient experience with care*, such as those collected via the Consumer Assessment of Healthcare Providers and System (CAHPS) surveys (developed by the Agency for Healthcare Research and Quality, or AHRQ). The various CAHPS measures assess experience with care in a variety of care settings and with various types of care. As with PRO-PMs, the informant for such measures is not necessarily the care recipient. For example, experience of care surveys for the hospice setting often rely on the family caregiver as the informant, while such surveys focusing on care provided to children may obtain data from parents. Another common focus of instrument-based measures includes those that elicit information on *health behaviors*.

Examples of concepts that could be assessed via instrument-based measures include:

- Satisfaction with care
- Extent of involvement in healthcare decision-making
- Feeling heard and respected during a healthcare episode

Many stakeholders view ***composite measures*** as another type of healthcare performance measure. Composite measures combine two or more component performance measures, each of which individually reflects some aspect of care, into a single performance measure with a single result. The component measures within a composite may be structure, process, or outcome measures, or any combination of these. The selection of component measures included in a composite measure, and the ways these are combined within the composite, will depend on the care construct the composite is attempting to assess. Examples of methods of aggregating components within a composite measure include using an arithmetic mean or weighted means. While many stakeholders like the idea of composite measures to summarize performance across a variety of individual component measures, composite measures do not reduce data collection burden and may pose challenges for interpretation and use, particularly for improvement (as opposed to accountability).

Anatomy of a Measure

While not all measures are constructed in the same way, many include the following components:

- **Numerator**: the number of patients (or events) who meet the focus (or intent) of the measure
- **Initial population**: the number of patients (or events) who meet the measure's target (or conceptual) population
- **Exclusions**: the number of patients (or events) from the measure's initial population who should dropped from the measure
- **Exceptions**: the number of patients (or events) from the measure's initial population who should be conditionally dropped from the measure
- **Denominator**: the total number of patients (or events) from the measure's initial population who are included in the measure after exclusions and exceptions have been applied

Using the measure "Intravesical Bacillus-Calmette Guerin for Non-muscle Invasive Bladder Cancer" as an example:

- The numerator includes those patients for whom BCG is initiated within six months of the initial bladder cancer staging.
- The initial population includes all patients initially diagnosed with T1, Tis or high grade Ta non-muscle invasive bladder cancer with a qualified encounter during the measurement period.
- The exclusions include patients who are immunosuppressed, have active tuberculosis, have mixed histology urothelial cell carcinoma, and those who undergo cystectomy, chemotherapy or radiation within six months of bladder cancer staging.
- The exceptions include those patients who did not receive BCG in the measure timeframe due to the unavailability of BCG.
- The denominator includes all patients from the initial population minus all exclusions and exceptions.

The result for this measure, for a particular urologist, is as follows:

$$Result = \left(\frac{numerator}{initial\ patient\ populaton - exclusions - exceptions}\right) * 100 = \left(\frac{numerator}{denominator}\right) * 100$$

Note that, by convention, implementers often mislabel the initial population by referring to it as the measure denominator, when, in fact, the actual denominator may be a subset of the initial population if exclusions and/or exceptions are applied. Of course, the initial population is the same as the denominator when there are no exclusions or exceptions.

Data Sources for Healthcare Performance Measures

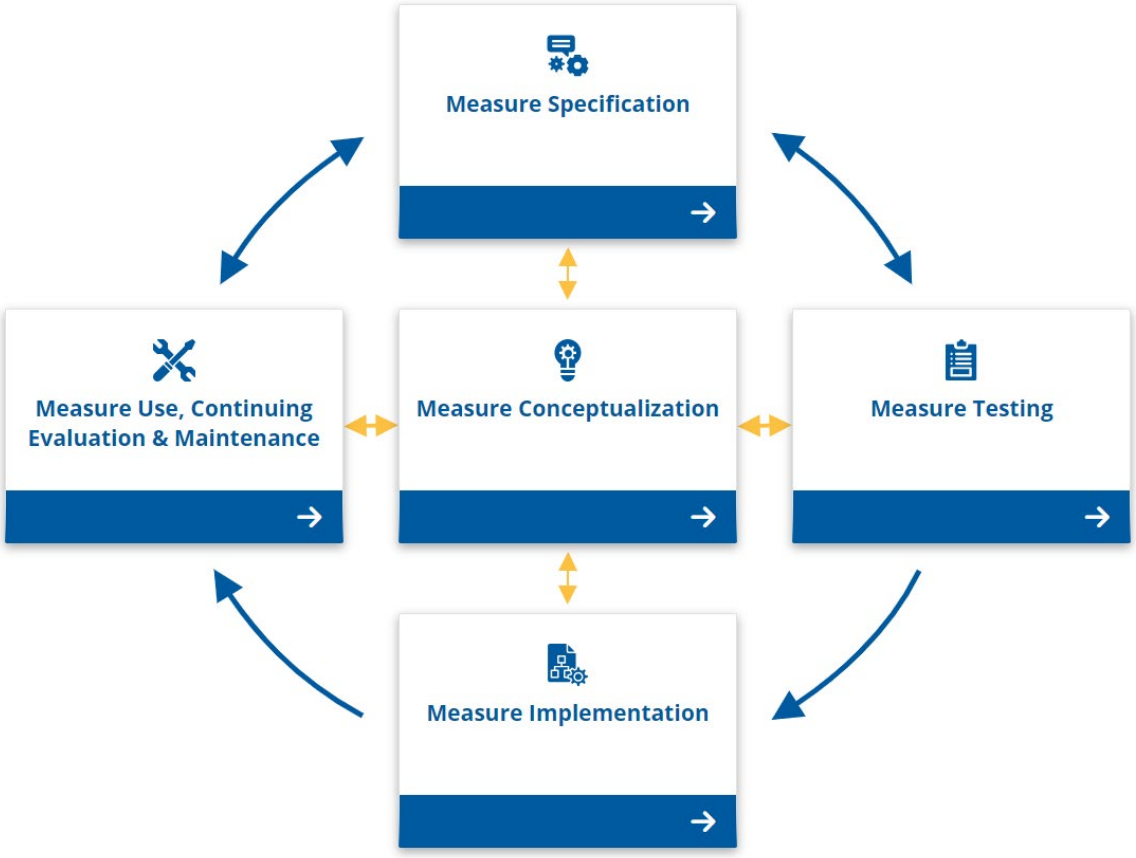
Many sources of data can be used in healthcare performance measures. Some of the most common include:

- Paper medical records
- Electronic health records (EHRs)
- Other electronic clinical data (e.g., pharmacy, labs, imaging data)
- Electronic assessments (e.g., MDS—assessment data from nursing homes; OASIS—assessment data from home health agencies)
- Administrative claims (e.g., insurance claims) and other types of administrative data
- Clinical data registries (e.g., AQUA, SEER)
- Patient reports (e.g., from surveys such as CAHPS)

Each of these data sources have pros and cons in terms of their feasibility and utility in healthcare performance measurement. For example, the data collection and measure reporting burden is very high for measures relying on paper medical records. Conversely, measures relying on data collected through electronic means may be easier to develop, test, calculate, and/or report. However, such measures may be limited because of the information available in such sources, particularly if the data are collected primarily for other purposes (e.g., administrative claims data collected primarily for payment purposes). Moreover, much electronically available data may not be easily usable for measurement (e.g., important information in EHRs available only in notes or imaging data available only via pdf files). Relatedly, many health data sources and systems are not interoperable, and thus combining data from various sources for the purposes of healthcare performance measurement may be difficult or impossible.

Measure Lifecycle

The measure lifecycle includes five key stages: conceptualization, specification, testing, implementation, and maintenance and continued use. CMS depicts the stages of the measure lifecycle as follows:



As with the classifications of measure categories and types, there is "squishiness" in this schema of the measure lifecycle (e.g., in the "location" of key activities within the stages or the sequencing of stages and other associated activities). Importantly, the various stages are not necessarily sequential and often are iterative, as activities and information in one stage may inform or even compel activities, information, and decisions in other stages. With de novo measure development, the initial stage is conceptualization. However, not all measure development is de novo.

CONCEPTUALIZATION

The conceptualization stage involves information gathering and idea exploration. Measure developers should consider the goals of the development effort (e.g., to improve accessibility of care) and the intended use of the measure (e.g., for internal improvement efforts or a specific type of accountability program). Such considerations inform the decision regarding the entity that will be held accountable for the measure (e.g., an individual clinician or clinician group, a hospital, etc.). Activities included in the conceptualization stage include reviewing the literature to understand the nature and magnitude of the problem of interest and obtain clinical evidence to support measure construction. Developers also should search for relevant existing measures to identify what gaps in measurement, if any, exist. Measure development is resource intensive, so care should be taken not to duplicate existing measures. Additionally, knowledge of relevant existing measures is important because developers can learn from previous development efforts by others.

The conceptualization stage also includes gathering feedback from experts in the clinical domain of interest, as well as from other relevant stakeholders who will have valuable perspectives that should be considered when developing a measure. Such stakeholders include, but are not limited to, those skilled in the technical aspects of development, patients and families who can provide insight on what they believe is important for measurement, and those who pay for the care that is being assessed. Such feedback should include input on data availability, data limitations, suggestions for clinical literature to support the measurement effort, and the costs and feasibility of both development and implementation, among many other potential topics.

The conceptualization stage ideally will conclude with one or more measure concepts that address the measurement goal of interest, or with the decision not to proceed with the measurement effort. Measure concepts should define the target population for the measure, the focus of the measure (usually, this is the measure numerator), and probable measure exclusions.

SPECIFICATION

Once measure concepts have been articulated, developers can begin creating the measure specifications. Measure specifications are the detailed instructions on how to calculate a measure. They should be **complete** and **unambiguous** so that the calculation of the measure result is consistent, even if done by different entities. Typically, measure specifications include narrative statements describing each component of the measure (e.g., target population, numerator, exclusions, etc.), detailed instructions on how to collect data for the various components of the measure, and instructions (in words, formulas, diagrams, or some combination) on how to aggregate the various components of the measure to get to the desired measure result.

Using "*timely receipt of BCG after bladder cancer diagnosis*" as an example measure concept, components that would be included in a measure specification include:

- The ***accountable entity.*** This is the entity, such as an individual clinician (a urologist) or clinician group (a urology practice) whose performance is being assessed by the measure.
- The ***data source*** used for the measure (e.g., data from a clinical data registry such as AQUA). Depending on the source of data, additional information may be included. For example, if data should be derived from a sample of data from a physician practice, the specifications must provide instructions on how the sample should be obtained (e.g., a random selection of 50 patients in a specified timeframe).
- ***What is being measured*** (i.e., timely receipt of BCG). The specifications must explicitly articulate what is meant by "timely" (e.g., within 6 months of diagnosis). They must also explicitly articulate how to identify receipt of BCG (e.g., specific CPT and HCPCS codes). For many measures, this will be the measure's numerator.
- ***What patients are targeted for inclusion in the measure*** (e.g., those initially diagnosed with T1, Tis, or high-grade Ta non-muscle invasive bladder cancer). The specifications must explicitly articulate how to identify these patients (e.g., specific ICD-10 codes and specific instructions on how to identify disease stage). The specifications must also state the timeframe for the patient population included in the measure (e.g., anyone diagnosed in 2023).
- Any of the initially targeted patients who **should not be included** in the measure. This may comprise those who are always excluded (i.e., **exclusions**, such as those who have undergone cystectomy, chemotherapy, or radiation within 6 months of bladder cancer staging) and/or those not included because to do so would unfairly penalize the accountable entity (i.e.,

exceptions, such as those who did not receive BCG due to shortage of the medication). Again, specific codes or instructions for identifying these patients must be articulated.

Depending on the complexity of the measure, additional information may be required. For example, outcome measures often are risk-adjusted to account for differences in patient characteristics at the start of care. For such measures, the specifications must include the factors included in the risk-adjustment model as well as their associated parameters. (Note that the considerations and processes involved in developing a risk-adjustment approach is beyond the scope of this Primer.) Similarly, measures of resource use/cost must articulate explicitly the approach used for attribution. Also, instrument-based measures, including PRO-PMs, must state which instrument(s) can be used for the measure, and potentially, which items in the instrument must be used.

Finally, as noted earlier, instructions on how to calculate the measure also should be included. Such instructions may be simple for basic process measures but would be much more complex for risk-adjusted outcome measures, cost, and instrument-based measures.

TESTING

Measure testing is empirical analysis conducted to demonstrate reliability and validity of a measure and its underlying data, in the context of its use. Many measurement scientists also view assessment of technical feasibility as an important part of testing. Typically, feasibility testing addresses issues related to data availability, cost, and burden of data collection (including impact on clinician workflow), barriers in calculating the measure or reporting its result (e.g., threats to patient confidentiality), and considerations of potential unintended consequences.

Often, testing is done concurrently with specification (e.g., a specification is drafted, then tested, which reveals some kind of problem; this drives modification of the draft specification, which, in turn, likely would require re-testing). Note that if testing reveals substantive problems with the measure as specified, the measure concept itself may need to be revisited.

Testing for both reliability and validity can be done for the data elements included in a measure as well as for the measure result. In the context of healthcare performance measures, reliability refers to the repeatability and precision of measurement, while validity refers to the accuracy or correctness of measurement (see table below). Often, initial testing focuses on the data elements (particularly on the assessment of feasibility and validity of the data used in the measure). Reliability and validity testing of measure results requires adequate sample sizes; as such, this level of testing may need to occur later in the measure lifecycle once enough data become available.

	Data Elements	**Measure Result**
Reliability	Consistency/repeatability of data collection	Precision of the result; ability to correctly distinguish performance
Validity	Accuracy of data used in the measure	Correctness of conclusions about care that are based on the measure results

Due diligence in ensuring reasonable reliability and validity of measure results and their underlying data helps to minimize waste of resources in collecting data, reporting results, or acting on results that

might be dubious. Ultimately, testing helps to reduce misinformation, misdirection, and unintended harmful consequences for patients.

Ideally, measures will be tested for both reliability and validity, for both the data elements and the measure results. Moreover, measures should be re-tested on a routine basis to ensure they continue to function as intended over time. Approaches and methods used in testing will vary depending on the goals of testing, the "level" of testing (i.e., data element or measure result), the type of and complexity of the measure, and the availability of data and other resources (e.g., time, expertise, money).

IMPLEMENTATION

The implementation stage encompasses the activities involved in putting a measure into use once it has been developed. For AUA-developed measures, this may be as simple as working with our vendors to collect and code the relevant data in the AQUA Registry, calculate the measure results, and display those results through the AQUA dashboard. However, the AUA must seek approval from CMS for a measure to be available for reporting in the CMS Merit-Based Incentive Payment System (MIPS) program. If offering a measure for MIPS reporting only via AQUA (a CMS Qualified Clinical Data Registry), the AUA must request this approval on an annual basis. This includes summarizing the clinical evidence and rationale for the measure, describing the measure specifications, providing current performance results, and documenting reliability and validity testing results. If making a measure available for MIPS reporting to any clinician regardless of the data collection mechanism (i.e., outside of AQUA), it must undergo the CMS rule-making process, which is both time- and resource-intensive.

MAINTENANCE AND CONTINUED USE

At the AUA, measure maintenance activities, along with other internal and external priorities, inform decisions about whether and how to keep measures relevant in an ever changing environment. Measure maintenance, which is conducted on an annual basis, involves:

- Reviewing the clinical evidence underlying the measure to ensure it still supports the measure as specified
- Assessing current performance to ensure there is still a deficiency that can be impacted by the measure
- Conducting additional testing or other analyses as needed, and
- Updating the measure specifications as needed.

For example, if CMS refuses approval for a measure to be used in the MIPS program, the AUA might retire the measure altogether (meaning it would no longer be calculated or maintained) or it might be kept in AQUA, but only for use in local quality improvement efforts or for research or historical purposes.

Beyond the Basics

This section includes a brief introduction to several more advanced topics in healthcare performance measurement. These topics are included primarily to underscore the complexity of the measurement enterprise, and thus should not be considered exhaustive.

Characteristics of Good Measures

How do you know if a measure is a "good" measure? This question is relevant for both developers and users of measures. For 20 years, the National Quality Forum (NQF) served as the consensus-based entity in the U.S. that evaluated measures using standardized criteria and conferred NQF endorsement status on those deemed to best suited to achieve the goal of high quality, efficient healthcare for individuals or populations. NQF's measure evaluation criteria were developed to reflect desirable characteristics of performance measures. ***Thus, measures that most closely align with NQF's endorsement criteria can be considered good measures***. Importantly, these criteria evolved over time, reflecting changes in the performance measurement enterprise (e.g., additional criteria for evaluating cost measures, composite measures, and PRO-PMs). Although NQF no longer endorses measures, the Partnership for Quality Measurement (PQM), the entity that currently evaluates measures for endorsement, uses criteria based on those developed by NQF.

NQF's major endorsement criteria included:

- **Importance to measure and report**: This criterion recognized that there are myriad important structures, processes, and outcomes in healthcare, but not all "rise to the level" required for measure development, implementation, and reporting of measure results. NQF's goal for this criterion was to endorse those measures focusing on aspects of care with the greatest potential of driving improvements.

  This first major criterion had two "non-negotiable" subcriteria. First, measures should be ***evidence-based*** (notably, this meant empirical evidence, not expert opinion). Second, there must be a ***quality (or access or cost) problem*** that a measure can address. This could be demonstrated either by observing variation in performance across accountable entities, less-than-optimal performance overall, or variation in performance for population subgroups (i.e., disparities).

  Of note, NQF viewed the Importance criterion (and these subcriteria) as essential and paramount. Measures that did not meet this criterion were not endorsed, regardless of how well they were aligned with the remaining endorsement criteria.

- **Scientific acceptability of measure properties**: NQF's goal for this criterion was to endorse those measures that allowed stakeholders to make sound conclusions based on measure results. Underlying "non-negotiable" subcriteria included measure ***reliability*** and ***validity***. Reliability requires complete and unambiguous specifications and can be demonstrated via testing. Validity also can be demonstrated via testing, but must also consider potential threats such as missing data, the choice of a risk-adjustment approach, etc. Consideration of reliability and validity can help to address questions such as:
  - Are the specifications clear enough so that everyone will calculate the measure in the same way?
  - Is the variation in measure results between providers primarily due to real differences? Or is it because there is a lot of "noise" in the measurement?
  - Does the measure actually assess what it is intended to assess?
  - Do the results of the measure allow for correct conclusions about care delivery?

As with the first major criterion, NQF viewed scientific acceptability as essential; measures that did not meet this criterion were not endorsed, regardless of how well they aligned with the remaining endorsement criteria.

- **Feasibility**: NQF's goal for this criterion was to endorse those measures that minimize the burden of data collection and measure implementation to the extent possible. Thus, measures utilizing electronic data sources without a lot of "extra" work outside of routine clinical processes were prioritized for endorsement. NQF recognized there is a "spectrum" of feasibility; thus, feasibility of data collection and implementation in every possible scenario was neither expected nor required. This is significant because there can be a tendency to focus on things that are easy to measure (although those may not be the most critical or useful things to measure).
- **Usability and Use**: NQF's goal for this criterion was to endorse those measures that could be used both for quality improvement and for accountability. NQF used this criterion to promote transparency (particularly via public reporting of measure results) and fairness (i.e., those held accountable for a measure should have the ability to provide feedback that would be taken seriously and be given assistance in interpreting and using measure results to improve care). Most importantly, however, this criterion evaluated the ability of measures to achieve high-quality care delivery by focusing on actual ***improvement*** (either in measure results or in downstream outcomes) and in ensuring the benefits of a particular measure ***outweighed any unintended negative consequences*** to individuals or populations due to the measure. Examples of measures with unintended negative consequences include measures that promote "cherry picking" of patients or measures that discourage screening or counseling.
- **Comparison of related or competing measures**: NQF's goal for this criterion was to endorse the superior measure when duplicate measures were available and to align measure specifications to the extent possible when similar measures were available. This criterion recognized the potential burden and confusion resulting from duplicative or similar-but-different measures.

Preferred Measure Types

Each measure type has pros and cons. For example, structure and process measures are relatively easy to develop. In contrast, outcome measures are not as easy to develop due to potential data limitations and the likely need for risk adjustment. Clinicians often prefer process measures because they can control or strongly influence workflows related to the focus of a measure; moreover, using process measures for quality improvement often is straightforward (i.e., one knows exactly what process(es) to intervene and improve on). In contrast, payors, policymakers, and patients often prefer outcome measures, particularly for accountability programs. For example, CMS has explicitly stated its preference for outcome measures for the MIPS program. Similarly, as part of its endorsement activities, NQF publicly stated its preference for health outcome measures and PRO-PMs, followed by measures of intermediate outcomes most closely linked to outcomes, and then by measures of processes or structures most closely linked to outcomes. NQF supported its preference for outcome measures in part because outcomes are what drive patients to seek care and drive clinicians to provide care (e.g., symptom relief, improved function, survival). Outcome measures also encourage a systems approach in care delivery and enable flexibility for improvement efforts at the local level (e.g., different groups may identify and intervene on various care processes or systems to improve rates of surgical complications).

Currently, the AUA does not have a stated preference for particular measure types. Instead, we believe there is a place for all types of measures, and ideally, we will develop and/or implement a variety of measure types.

Risk-Adjustment

Risk adjustment (also known as case-mix adjustment) is used to "level the playing field" by controlling for patient (or other relevant) factors present at the start of care so that accountable entities can be compared based on the care they provide, not on the characteristics of the patients they see. Risk adjustment is expected for outcome and cost measures, although it may not always be required. It is usually accomplished via a statistical modeling approach, which is itself both an art and a science, requiring a high level of data and statistical proficiency. Determining which risk factors should be considered—and ultimately included—in a risk-adjustment approach is not straightforward, and conceptual, statistical, and even philosophical issues come into play. For example, stakeholders have opposing views on whether social risk factors (e.g., income, education, marital status, etc.) should be included. Some think these factors should be included to maximize fairness for those being measured. Others think they should not be included because of the fear that doing so "gives permission" to provide lower-quality care to socially vulnerable subgroups. Data availability can be a critical limitation for effective risk adjustment (e.g., relatively large sample sizes are required; critical data elements may not be available or may have a large proportion of missing values).

Low Case-Volume

Low case-volume presents a significant measurement challenge that is particularly salient for small and/or rural providers and practices. However, it can also be a challenge for condition- or procedure-specific measures, particularly when they focus on relatively uncommon conditions (e.g., bladder cancer). For the most part, the low case-volume problem is one of not having enough patients in the measure denominator. It is of particular import for measures used in accountability applications, but also affects ability to gauge improvement. Low case-volume primarily affects the reliability of measure results, although it can also affect validity (e.g., through its effect on the adequacy of risk adjustment). A number of approaches for measure development and/or measure selection can help alleviate the low case-volume problem. Examples include reconsidering how a measure is calculated (e.g., using continuous variables, ratio measures, judicious use of exclusions), incorporating additional data (e.g., pooling data over longer time periods or more providers), or using more complex statistical and computational methods in measure calculations (e.g., partial pooling of data across both time and providers).

Fundamental Tensions in Healthcare Performance Measurement

Healthcare performance measurement—both development and use--must balance competing desires of disparate stakeholders. Some of the most fundamental tensions in measurement include:

- Intended use: Whether to focus on a small number of outcome measures for use in accountability applications versus developing and using measures that focus on specific structures or processes to guide improvement
- Accountable entity: Whether to measure at a system level versus measuring performance of individual clinicians
- Measurement burden: Whether to prioritize burden for clinicians in data collection and

reporting versus desire by consumers and purchasers for more (or more complex) measures to aid in payment decisions or consumer choice

- Variety: Whether to offer a small "core set" of measures versus offering a larger "menu" of measures that meet the needs of different specialties and settings

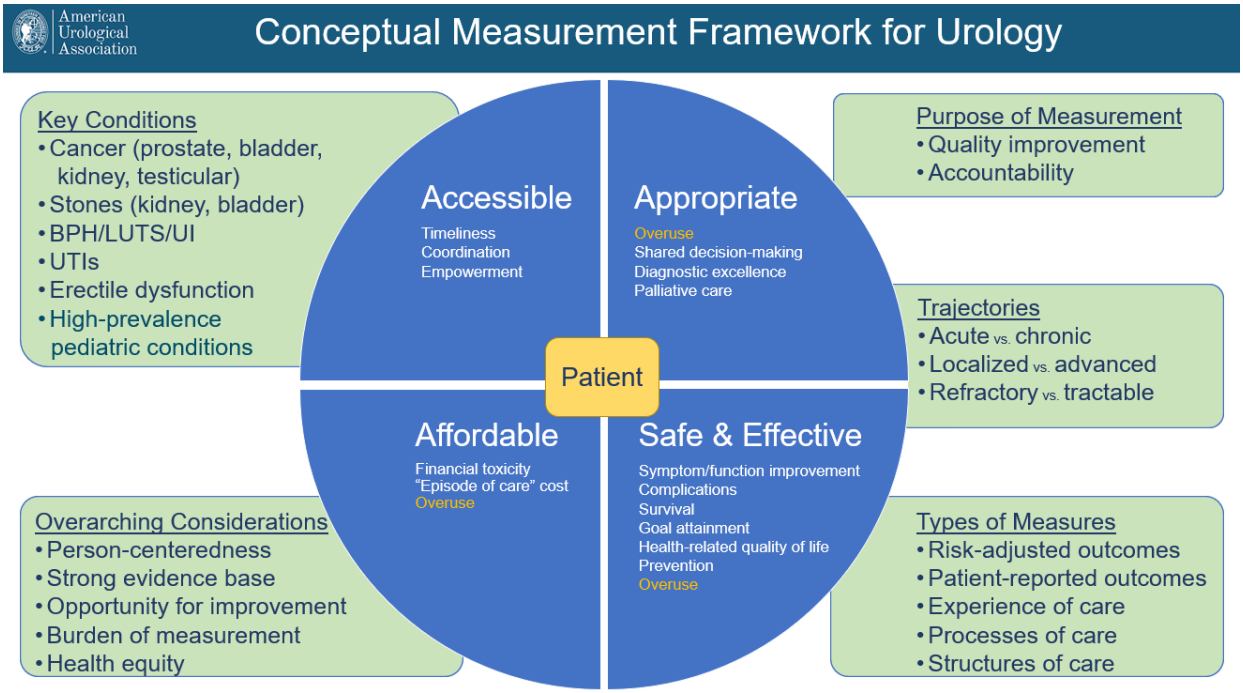
Note that the above are inter-related, and the list is not exhaustive.

Future Directions

Healthcare performance measurement is a constantly evolving enterprise, both in general and for the AUA more specifically. For example, the desire for value-based care in the U.S. directly impacts the design of measurement programs, which in turn drives the number, types, and even topics of measures included in such programs. The movement toward data and system interoperability is changing the way measures are (and will be) developed and implemented. In addition, measures are becoming more complex and sophisticated, as illustrated by the increasing numbers of risk-adjusted cost and outcome measures and PRO-PMs. Furthermore, advanced computational techniques such as natural language processing and generative AI almost assuredly will allow measurement of ever more important concepts that are infeasible to measure currently.

The AUA's measurement activities in recent years have focused on strategy and maintenance rather than de novo development. For example, we have assembled at 25+ member Measure Evaluation Panel (MEP) to aid in measure development, maintenance, and evaluation. This large group brings a wealth of needed perspectives and expertise to inform our measurement activities. Also, in 2021, AUA's QIPS Committee reviewed and approved a conceptual measurement framework to help guide AUA's measurement efforts over the next 5-10 years. This framework was updated in 2024 (see Appendix A) and will be reviewed periodically in the future (and revised as needed) to ensure its continued relevance for the field. Additionally, with input from MEP members, AUA staff have worked to revise AUA-develped measures, as needed, to respond to feedback from users, conform to the most recent clincial guidelines, align with current coding schemas, and enhance clarity and precision. Looking ahead, we plan to develop or otherwise enhance measurement of disparities in urologic care and develop measures that focus on additional urologic conditions and/or relevant subpopulations.

Appendix A: Conceptual Measurement Framework for Urological Conditions



Appendix B: Additional Measurement Resources

CMS

Quality Measures: How they are developed, used, and maintained

CMS Quality Measure Development Plan (relevant to the Quality Payment Program)

CMS Measures Management System Blueprint (and supplemental materials)

CMS Meaningful Measures Initiative

Quality Payment Program

CMS Measure Inventory

National Quality Forum

The ABCs of Measurement

Partnership for Quality Measurement

Repository Measure Database

Endorsement & Maintenance (E&M) Guidebook