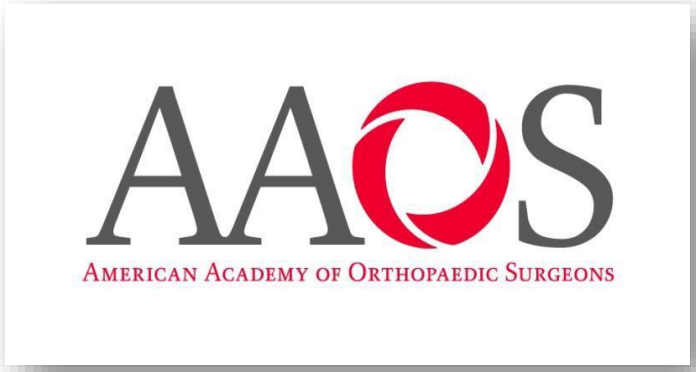


Quality Performance Measures

Methodology Document

AAOS Clinical Quality and Value

| Updated 2018

AAOS
American Academy of Orthopaedic Surgeons
American Association of Orthopaedic Surgeons

Table of Contents

Introduction

In an attempt by government and private industry to countervail unnecessary healthcare expenses, clinical quality and value have become the focus of research, administration, policy, legislation, and quality assurance and improvement initiatives. Healthcare has turned to value-based-care payment models, evidence-based research, and other forms of quality-related initiatives, and has begun to abandon or minimize the use of traditional payment models like fee-for-service. According to the Centers for Medicare & Medicaid Services (CMS), national healthcare expenditures currently comprise 17.9% of US Gross Domestic Product (GDP) (2016), with 'hospital care' and 'physician & clinical services' – two healthcare subcategories, accounting for 50% of each healthcare share[1]. This equates to a current cost of $10,348 per person-year and a total annual cost of $3.3 trillion in 2016[1]. According to the U.S. Department of Commerce, Bureau of Economic Analysis, in 2017 efficiency, productivity, and profitability of an economic enterprise (as represented by 'value-added', expressed as a percent of GDP) was 7.3%; by comparison, real estate was 13.4%, and professional and business services was 12.1%[2], evidence that the U.S. is experiencing an unfavorable cost/return imbalance in the healthcare sector. Put another way, we spend a lot, but do not realize a value commensurate with our spending. Furthermore, spending has increased in nearly every healthcare domain, from Medicare, Medicaid, private insurance, prescription drugs, and out-of-pocket expenses[1]. These issues are further compounded by an aging population. In 2011, the US Department of Health and Human Services, authored the initial *2011 Report to Congress: National Strategy for Quality Improvement in Health Care* report, advocating for better and more affordable care[3]. The report lists the National Quality Strategy's aims and priorities[3,4]:

- Safer care;
- Effective care coordination;
- Person- and family-centered care;
- Prevention and treatment of leading causes of mortality;
- Supporting better health in communities; and,
- Making care more affordable.

Healthcare performance measurement is one requirement of the major bipartisan healthcare legislation – Medicare Access and Children's Health Insurance Program (CHIP) Reauthorization Act of 2015 (MACRA), and its Quality Payment Program (QPP) and Merit Based Incentive Payments System (MIPS), for all eligible providers seeking payment for their Medicare patients. This legislation, among other factors, can result in notable payment adjustments. There has also been increased focus from payers and regulatory agencies to quantify and qualify the care rendered to patients. At a high level, healthcare performance measures are indicators of the quality of care rendered by clinicians to patients. The National Quality Forum (NQF), a CMS-funded measure endorsing organization, describes performance measures as a method for calculating whether and how often a healthcare entity does what it should. Measures are based on scientific evidence about processes, outcomes, perceptions, or systems that relate to high quality care[5]. Performance measures, simply put, allow practitioners to measure clinical care delivered, because one cannot change or improve things that one cannot measure.

Goals for Measure Development

The AAOS believes that the primary purpose of quality measurement is to identify opportunities to measure and thus improve patient care and other health-related outcomes. Guiding principles for measure development include[6]:

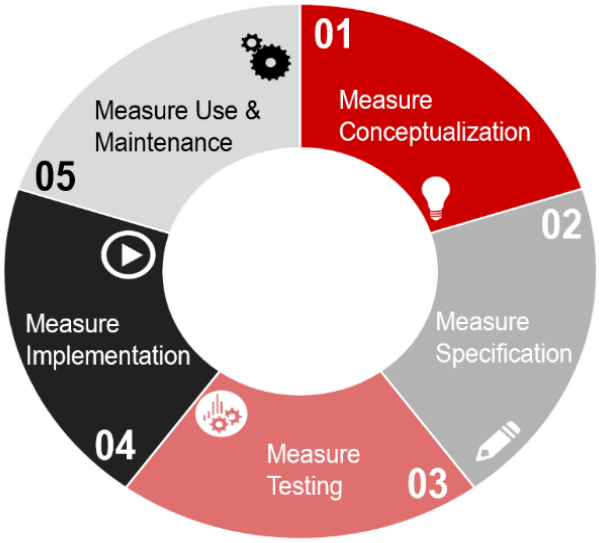
- Develop measures of high impact and importance for orthopaedic patients, differentiating prevalence, severity, and/or functional status;
- Develop measures that address a performance gap where there is known variation in clinical actions or the outcome(s) of interest;
- Develop measures that are supported by the best available evidence;
- Develop measures that are feasible; for example, data collection and report generation will not cause undue provider, patient, or caregiver burden;
- Develop well-defined measure specifications that incorporate broad stakeholder input;
- Develop measures that are statistically valid and reliable, ensuring that differences in performance scores reflects differences in clinical quality; and,
- Implement the final measures into CMS value-based payment programs, such as the QPP.

Quality measure development remains a focus for AAOS Clinical Quality and Value staff, to continually meet the evolving needs of its members, help its members meet MACRA and MIPS requirements, and to provide members with information that informs clinical decision making.

Measure Development Lifecycle

AAOS performance measures are developed using a rigorous, standardized process, commensurate with CMS, NQF and Health Level Seven International (HL7) standards. At the AAOS, measures are developed by physician-led work groups and supported by dedicated staff within the Department of Clinical Quality and Value. Figure 1 illustrates the five phases in the AAOS measure development lifecycle. Although the lifecycle shows each phase as a discrete activity, the measure lifecycle is dynamic. Some phases may overlap or take place concurrently or result in feedback with earlier phases.

Figure 1. Performance Measure Lifecycle.



Quality Measurement Overview

Anatomy of a Performance Measure

In most cases, at its basic level, a performance measure is a ratio. The denominator represents the number of eligible cases, less any exclusions or exceptions, and the numerator represents the number of instances the clinical action of interest was performed. It is helpful to note that the denominator is often derived from, and sometimes equal to, an initial population; this initial population is the broadest grouping (e.g. all patients age 16+ with a specified diagnosis)[6]. The initial population can be reduced to a denominator (e.g. all initial population patients that underwent surgical treatment), and then to a denominator with exclusions and exceptions removed. Figures 2 and 3 below visually depict the anatomy of a performance measure from an arithmetical perspective.

Figure 2. Anatomy of a Performance Measure.

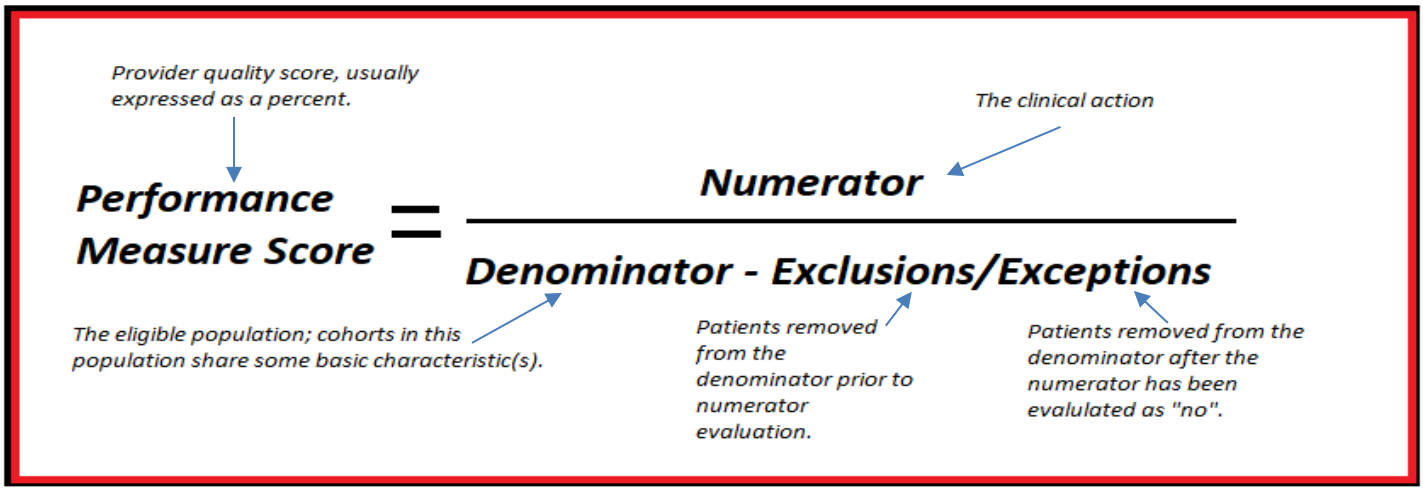
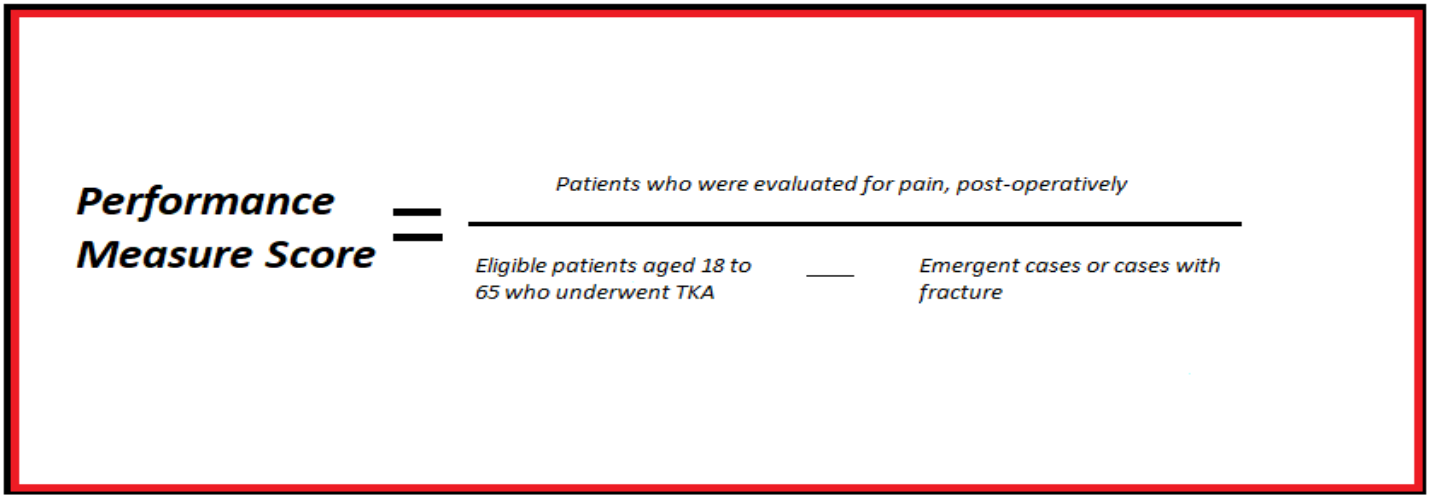


Figure 3. Performance Measure Example.



The Measure Developer needs to find a balance between power and specificity, including all relevant patients to ensure a large enough denominator for analyses, while also excluding or excepting patients for whom the denominator criteria are not appropriate or aligned with the measure's intent.

Denominator exclusions are those cases that are removed immediately from the denominator, before numerator evaluation, whereas denominator exceptions are removed from the denominator only after the numerator was properly assessed as "no".

The numerator can be viewed as a subset of the denominator, containing those eligible cases in which the clinical action of interest was performed. Thus, in Figure 3, the numerator patients would be those denominator patients (non-emergent, TKA patients) who received a post-operative pain evaluation. The numerator is the clinical action of interest.

Measure Types

There are several types of performance measures. Table 1 outlines the most common types of performance measures, their definitions, and examples[6].

Table 1. Types of Healthcare Performance Measures.

Type of Measure	Definition	Example
Structural	Structure of care is a feature of a healthcare organization or clinician related to the capacity to provide high-quality healthcare. Structure measures are supported by evidence that an association exists between the measure and one of the other CQM domains.*	The percentage of providers who have a system in place to track and follow patient falls.
Process	A process of care is a healthcare-related activity performed for, on behalf of, or by a patient. Process measures are supported by evidence that the clinical process—that is the focus of the measure—has led to improved outcomes. These measures are generally calculated using patients eligible for a service in the denominator, and the patients who either do or do not receive the service in the numerator.*	The percentage of women aged 50–85 years who have sustained a fracture and who either underwent a bone mineral density test or received a prescription for a drug to treat osteoporosis.
Intermediate Outcome	Often the intermediate step toward an outcome. Sometimes it is more appropriate to measure an intermediary, as opposed to an outcome.	The percentage of eligible providers who received fall prevention training; here, patient falls would be an outcome, but provider education would be an intermediate outcome.
Outcome	An outcome of care is a health state of a patient resulting from healthcare. Outcome measures are supported by evidence that the measure has been used to detect the impact of one or more clinical interventions. Measures in this domain are attributable to antecedent healthcare and should include provisions for risk adjustment.*	The percentage of surgical site infections occurring within 30 days after the surgical procedure. CMS is currently promoting the development of outcome measures, among others, as they are often patient-centered, more straightforward, and can be risk- and SES-adjusted.

Patient-Reported Outcome Performance Measures	Assesses patients' perspectives, regarding their care, including their assessment of any resulting change in their health, positive or negative.	Percentage of patients with >20 points difference in Hip Disability and Osteoarthritis Outcome Score (HOOS), pre- and post-operative.
Composite	Combines the results of two or more component performance measures, each of which individually reflects quality of care, into a single quality measure with a single score, to provide a more concise picture of quality care. Composite measures can simplify and summarize a large number of measures or indicators into a more succinct measurement.	Measuring in-hospital mortality indicators for select orthopaedic conditions.
Misuse/Overuse/Access	Access to care is the attainment of timely and appropriate healthcare by patients or enrollees of a healthcare organization or clinician. Access measures are supported by evidence that an association exists between the measure and the outcomes of or satisfaction with care.*	Percentage of providers who ordered an MRI for low back pain. Here, it is assumed that the MRI is not indicated as a first-line treatment, so we would be looking at overuse of MRIs for the study population.

*Definitions taken from the CMS Blueprint for the Measures Management System, V 13.0 (2017), Table 3, Section 4.

Outcome Measurement

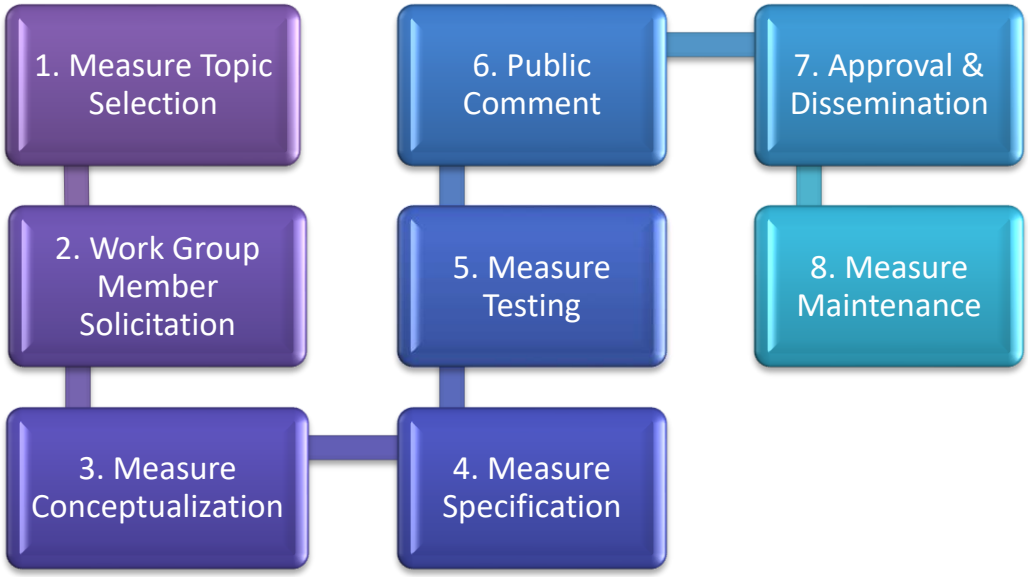
There has been strong encouragement from CMS and private payers toward developing outcome measures because process measures are not always a perfectly mapped surrogate for the desired outcome; in the example from Table 1 above, fall prevention education for providers does not necessarily mean reduced number of falls for patients, nor is education the only way to achieve a desired outcome. Thus, measuring outcomes may be a preferred approach over measuring processes. Outcomes measurement is not new in healthcare, medicine, or public health, but within the measurement science arena, outcome measurement is still new and developing.

Outcome measurement can also be supplemented with risk adjustment, a statistical method for controlling factors that are known to influence the relationship between the predictor and outcome. For instance, if the outcome of interest is surgical readmission rate, we may want to control for case severity, present on arrival conditions, or hospital resources (i.e. access to advanced imaging). Other statistical methods, such as risk standardization or stratification, can also be utilized. These methods become helpful in instances when the outcome of interest is influenced by a multitude of factors, often out of the control of the provider (i.e. surgical case severity or pre-existing conditions), and thus "leveling the playing field" by teasing out true differences in quality.

Measure Development Process

The measure development process can take as much as years to develop a measure, often depending on measure complexity, data acquisition, CMS or NQF measure review cycles, but typically includes eight phases (see Figure 4).

Figure 4. Performance Measure Development Process.



Topic Selection

The first step is an ongoing "call for measure topics" that is widely announced and publicized through multiple mediums including the AAOS website. Clinical topic suggestions for measure development can originate internally within the AAOS membership and/or externally. For each topic under consideration, the AAOS completes an environmental scan to determine the amount of concurrent research efforts to avoid duplicative efforts; the environmental scan also helps determine gaps in care and identifies opportunities to harmonize with related measures.

Work Group Solicitation

Purpose: After a topic has been chosen and a project timeline has been established, the measure development work group (e.g. the technical/clinical expert panel) must be convened. The purpose of the work group is to guide and direct the process of transforming a measure topic to a measure concept, and then measure specifications that can be tested and implemented.

Structure: Each project looks to convene a multidisciplinary panel of experts in the selected topic area for the measure development work group. Work group members are selected based on their expertise, experience, diversity of perspective, and training. Relevant medical specialty groups are considered and frequently invited to nominate members to serve on measure development work groups. Work group members include clinicians, quality improvement experts, methodologists, and statisticians. Work groups consist of approximately 8-12 members, including an Oversight Chair and Work Group Chair. The specific roles and responsibilities of each can be found in Appendix A. Note that an AAOS staff member serves as the Lead Methodologist/Project Manager for each work group.

Conflict of Interest: The AAOS is committed to producing non-biased and clinically important quality measures and goes to great lengths to ensure the integrity of the development process. As such, bias is addressed beginning with the selection of Clinical Practice Guideline (CPG) and Systematic Review (SR) work group members. The AAOS conflict of interest (COI)

policy is strictly enforced and all work group applicants must complete an enhanced disclosure as part of the application process, which is then evaluated for potential relevant financial COI. Applicants with financial COI related to the CPG or SR topic cannot participate if the conflict occurred within one year of the start date of the measure development or if an immediate family member has, or has had, a relevant financial conflict. Additionally, all development group members sign an attestation form agreeing to remain free of relevant financial conflicts for one year following the publication of the measures. Financially conflicted members may, however, be invited to either (1) divest themselves of relevant financial COIs or (2) remain involved with the project as a non-voting consultant.

Measure Conceptualization

Whereas a measure topic includes the clinical area or condition, the measure concept includes both the clinical area/condition and the aspect of care to be measured. An example of a measure concept might include implementing a multimodal approach for pain management in total knee arthroplasty. Measure conceptualization begins with reviewing the relevant evidence, such as the strong and moderate-strength recommendations from AAOS CPGs. Next, the work group prioritizes measure concepts based on related and competing measures, measure feasibility, impact on improving patient care and importance, usability, the potential for unintended consequences, industry trends, and the potential to reduce inequality in healthcare delivery. A useful tool to utilize during this phase is the NQF criteria matrix. The criteria matrix helps developers evaluate and prioritize measure concepts. See Appendix B for an example. If there are related or competing measures, work groups may want to harmonize measures, or identify whether existing measures can be adopted or retooled to fit the desired purpose of a new measure. During measure conceptualization, the work group considers:

- What important processes or outcomes need to be improved?
- Do the processes relate to improved outcomes?
- Can the outcomes be influenced?
- Can the processes or outcomes be measured?
- What will the measures be used for?

More on Measure Types – Qualified Clinical Data Registry PMs

As described in the "Measure Types" section above, there are different types of measures (i.e. process, outcome, patient-experience), so named by the aspect of care they measure. Measures can be further named in accordance with the data source they require. For instance, there are registry, QCDR PMs, eMeasures, eCQMs, claims, and paper-based measures. Each of these measures has specific development, specification, testing, and publishing requirements. The AAOS's emphasis on registry participation and status as a QCDR offer an opportunity for development and maintenance of impactful measures which can be frequently refined using clinical data and used to drive quality improvement for registry participants. Table 2 below describes the CMS requirements for the development of QCDR measures.

Table 2. CMS QCDR Measure Development Requirements.

Centers for Medicare & Medicaid Services – QCDR Measure Self-Nomination Requirements		
	Required	**Favored**

Administrative	Must be submitted during self-nomination period.	Supports quality improvement activities.
	Yearly re-nomination requirement.	
Development	Evidence-based; evidence ideally within last 3 years, or most recent evidence is used.	Addresses a measurement gap.
	High appeal, high face validity.	Numerator is meaningful and represents quality or clinical action, and not documentation.
	Development (i.e. concept formation) of measure should be complete.	Measure support the NQS priorities: patient safety, person/caregiver centered, outcome, care coordination, effective clinical care, community/population health, efficiency and cost reduction.
	Measure classification identified, datatype, data source,	CMS wants 'high-priority' measures: outcome, appropriate use, patient safety, efficiency, patient experience, care coordination, composite, and multi-strata measures.
Specification	Feasibility and implementation analysis complete.	Supports/favors PROM-PMs.
	Feasible, flexible, responsive.	Risk-adjusted measures.
	Not burdensome to report	
	QCDR Measures are not in the MIPS program, unless their MIPS (eCQM) version significantly differs from the QCDR version; do not need an eCQM version.	
Testing	Full specs developed, feasibility and implementation analysis completed.	Validity and reliability computed. Validation process provided.
	Current performance shows variation, or, current literature suggests variation; room for improvement in score, not topped out.	

Measure Specification

Healthcare performance measures need to be formatted in a manner that enables widespread, seamless implementation and use. This formatting is called specification. Performance measure specification is the process of defining healthcare clinical concepts using standardized terminologies and formats that are recognized by common healthcare technologies. These standardized formats also enable easier technological implementation, measure concept harmonization, interoperability, and information sharing. In their Blueprint, CMS refers to measure specification as the "how" and "where" of data expression and capture[6].

The Quality Data Model (QDM)[7], developed and maintained by CMS' partners – Office of the National Coordinator for HIT (ONC), MITRE and ESAC, Inc., is the gold standard model for standardizing clinical quality concepts (e.g. performance measures, eMeasures, and benchmarking analyses), so that stakeholders within healthcare quality can express, communicate, share, and exchange information effectively and accurately, using its standardized process. The current

QDM[7] (version 5.3), used in conjunction with the Measure Authoring Tool (MAT), currently expresses authored measure logic in Clinical Quality Language (CQL, version 2), a logic expression that replaced QDMs' expression in late 2017; CQL can better express complex clinical concepts, is human readable, and technology- and query-friendly, making it the new trial standard.

Table 2 below demonstrates how a performance measure transforms from clinical concept(s) to full technical specifications. Measure Developers may consider developing a Data Requirements Matrix, a technical array containing all required data and clinical concepts, in progressive granularity, according to the QDM data categorizations[6,7]. Once the Measure Developer completes the customized Data Requirements Matrix, that matrix, and the information it contains, can be authored in the MAT, and expressed using the new Clinical Quality Language (CQL)-based Health Quality Measure Format (HQMF), both HL7 standards. Here, the final product will be a draft eCQM. For both registry and QCDR measures, a Data Requirements Matrix can aid in measure feasibility analyses and implementation efforts. The AAOS follows CMS' and ONC-HIT's QDM, Version 5.3, when developing eMeasures. Figure 5 below depicts a Data Requirements Matrix Table.

Table 2. Specifications: Required Data Matrix[6].

Data Requirements Model & Specifications

Data Category	Datatype & Attribute	Code System	Value Set or Direct Reference Code
**High-level QDM category; there are currently 21 recognized categories.* **Value sets define the category.*	**Context of QDM category.* **Provides detail about the QDM element.* **Value sets do not define the datatype.*	**Coding and classification system; terminology, standard.*	**Set of codes and/or terms, described within the coding system.* **Value sets can contain other value sets to make value set groupings.*
--Individual characteristic --Encounter --Diagnosis --Medication	--Patient characteristic --Medication, Dispensed --Medication, ordered --Encounter, performed --Provider care goal	--Systematized Nomenclature of Medicine – Clinical Terms (SOMED-CT) --International Classification of Diseases (ICD) 9 or 10 --LOINC --Current Procedural Terminology (CPT)	--Age at time of encounter --Ethnicity --Date of discharge --BMI --CPT 25607 [direct reference]

Figure 5. Data Requirements Matrix.

Measure Component & Clinical Concepts	QDM Category	Datatype & Attribute	Value Set, Data Element Name, Variable Name	Standardized System	OID, Value Set, Direct Reference	Timing and Other Constraints	Notes
Supplemental Data Elements	Example						
<women>	individual characteristic	patient characteristic	gender	HL7	OID: 2.16.000000000000.	at measurement period	measure calls for stratification
Initial Population							
had CT performed	Procedure	procedure, performed	CT	CPT	CPT 74176, 74178	before start of measurement period	need CT before and during MP
Denominator							
equals IP	-	-	-	-	-	-	-
Denominator Exclusions							
hip fractures	condition/diagnosis/problem	diagnosis	prim_diag	SNOMED-CT, ICD-10-CM	ICD-10 S32.301A	overlaps the measurement period	must be primary diag, not secondary
Numerator							
medication reconciled	intervention	intervention, performed	med_recon	SNOMED-CT	direct ref	before discharge	result

AAOS
American Academy of Orthopaedic Surgeons
American Association of Orthopaedic Surgeons

Measure Formatting, Standards, and Considerations

Once the Measure Developer has derived draft measure concepts from the development work group efforts, and as the specification process is underway, there are several formatting standards and other considerations one should keep in mind. These additional formatting standards and considerations are described below in Table 3.

Table 3. Measure Component Formatting.

Component	Measure Component Formatting & Considerations
Measure Name	<focus population> who received/had <measure focus> Special considerations: Use the following statements for appropriate use measures*: <appropriate use of> < appropriate non-use of> <inappropriate use of> Example: THA patients who had physical therapy ordered.
Measure Description	<percentage, proportion, number> + <focus population> + <measure focus> Example: Percentage of THA patients who had post-surgical physical therapy ordered.
Initial Population	Broadest group of cases, who share key characteristics*.
Denominator Population	The population to be evaluated; the population to which the numerator applies to*.
Numerator Population	The clinical action, process, episode, or event that satisfies the measure intent*.
Denominator Exclusions	Cases removed from the denominator, prior to evaluating the numerator; these cases are removed immediately*.
Denominator Exceptions	Cases removed from the denominator, after evaluating the numerator as "no". Exclusions and exceptions further define the denominator, or study population*.
Timing	Specifically define episode or event, and timing constraints or intervals, if any. Example: Physical therapy ordered in the 30-day post-surgical period. In the example above, "PT" is the event, and the "30-day post-surgical period" is the timing; here, the timing period can be (1) discharge date + 30 days, or (2) procedure date + 30 days, or (3) PACU date + 30 days.

<table>
<tr><td></td><td>*Note, the Measure Developer should assess for a timing component for each data element, to ensure the proper data is pulled for measure calculation. More information regarding how to be aware of timing and proper data capture is provided in the section below.</td></tr>
<tr><td></td><td></td></tr>
<tr><td>Defining precise episodes, events, claims, actions</td><td>Consider the following clinical concepts below and note that each of these concepts may pose issues that can result in the incorrect data being captured.

<table>
<tr><th>Clinical Concept</th><th>Potential Concerns Warranting Address</th></tr>
<tr><td>“medication reconciliation”</td><td>Pre-surgical or post-surgical? Day of procedure, or day of discharge?</td></tr>
<tr><td>“consult encounter”</td><td>Pre-surgical consult or post-surgical consult? Orthopaedic consult, psychological/psychiatric consult, cardiology consult?</td></tr>
<tr><td>“EKG ordered post-discharge”</td><td>Which claim is needed? The first, the last, the one associated with a diagnosis (i.e. dyspnea)?</td></tr>
<tr><td>“antibiotics ordered”</td><td>Pre-surgical or post-surgical? Prophylactic antibiotics?</td></tr>
</table>

Clinical concepts are rightly concise, but the corresponding technical specifications should also be elaborate and specific.</td></tr>
<tr><td></td><td></td></tr>
<tr><td>Define Data Source</td><td>EHR, claims (electronic or paper-based), registry, QCDR.</td></tr>
<tr><td></td><td></td></tr>
<tr><td>Define Code Systems</td><td>The Measure Developer must determine what code systems capture their clinical concepts. Examples include:

➢ ICD-9-CM, ICD-10-CM;
➢ ICD-10-PCS;
➢ CPT;
➢ HCPCS [Healthcare Common Procedure Coding System];
➢ SNOMED CT [Systematized Nomenclature of Medicine Clinical Terms];
➢ LOINC [Logical Observation Identifiers, Names, and Codes];
➢ RxNorm, NDC;
➢ DSM [Diagnostic and Statistical Manual for of Mental Disorders];
➢ CDC codes for race and ethnicity;
➢ MS-DRGs [Medicare Severity-Diagnosis Related Groups];
➢ Value Sets;
➢ UCUM [Unified Code for Units of Measure];</td></tr>
</table>

Define Level of Analysis	The Measure Developer needs to keep in mind the level and analysis, for instance, provider-level, hospital-level, plan-level, region (i.e. MSA), and so on. Most performance measures, especially eCQMs that are part of MIPS, will require reliability testing that is typically done at the provider-level, and validity testing that is done at the data element-level. The level of analysis must be considered when obtaining testing data, as well. Development of eCQMs will differentiate episode-of-care or patient-based measurement.
Define Project Goals	The Measure Developer should note their intended project goals, namely, the final measure deliverables. Will the measures you are developing be submitted to the Measures Under Consideration (MUC) list, submitted as self-nominated QCDR measures, or used for benchmarking purposes? These questions should be answered early in the development process.
*Definitions taken from the CMS Blueprint for the Measures Management System, V 13.0 (2017).	

Measure Testing

Once the performance measure is fully specified, measure testing can be initiated. Measure testing ensures the measure is feasible and scientifically acceptable (reliable and valid). Measures that undergo proper testing will demonstrate these traits:

- ✓ Feasible – The measure can be implemented without undue burden, and reports can be automatically generated with calculated measure scores.
- ✓ Reliable – The measure will calculate a score that is repeatable; differences in scores mean differences in quality, and not instrument or random error.
- ✓ Valid – The measure measures what it is supposed to, is accurate, and scores derived from the measure calculation are indicative of true quality.

Measure Testing Plan

A measure test plan should be developed early in the testing stage. This plan outlines the steps to be taken for the testing portion of the project. These test plans can become very helpful when collaborating with other stakeholders, hospitals/medical centers, other departments within the developer's or practitioner's organization, IRBs/Review Committees, Executives, or other entities, as they provide excellent summaries of the testing methodology.

Scientific Acceptability (Alpha Testing)

Scientific acceptability, a type of alpha testing, includes empirical validity and reliability testing of the draft measure(s). Validity ensures the measure is measuring what it is intended to measure and that the measure is consistent with current clinical practice guidelines; reliability ensures the measure calculates with repeated precision. Figure 1 below visually depicts different scenarios of reliability and validity, with the target on the far right being the goal (high reliability and validity).

Figure 1. Reliability and Validity Scenarios.

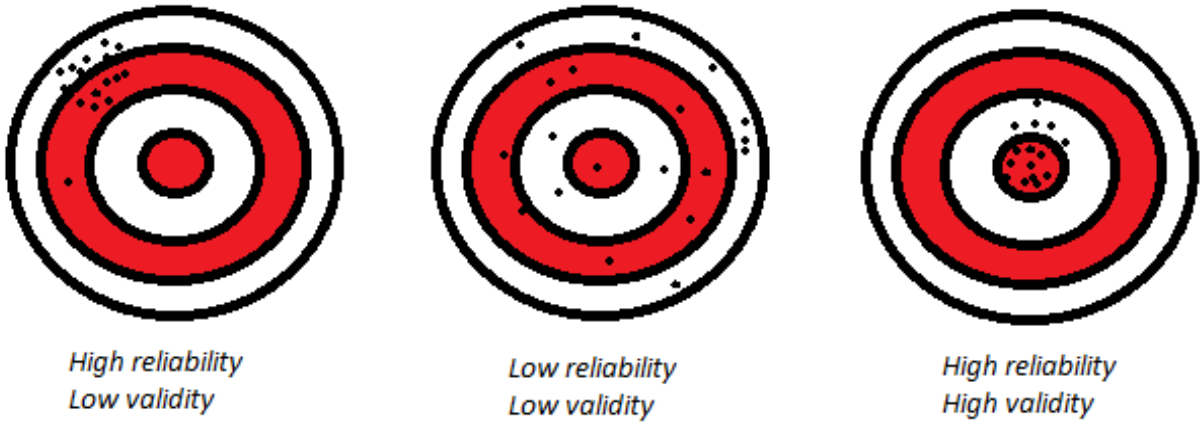


Table 3 below describes the most common types of scientific acceptability testing in measurement science and performance improvement.

Table 3. Scientific Acceptability Testing.

Type of Scientific Acceptability Testing	Description of Method
Signal-to-Noise Ratio Reliability	Repeatability and reproducibility of measurements. This method looks at the ratio of signal (i.e. numerator action) to noise (i.e. random variability and error), to determine real differences in quality. AAOS uses the Rand/Adams[8] method for computing signal-to-noise ratio testing. This method is currently accepted by NQF as meeting the reliability testing requirement, and the method is considered an industry standard. The general process of SnR analysis is as follows: 1) Structure and clean data appropriately for analyses. 2) Run a regression model (often a multi-level, non-linear, and/or mixed effects model) on the data; the model will estimate population parameters. 3) Enter these estimated population parameters into equations that calculate reliability, and compute. The generally accepted SnR threshold is 0.70 or greater[8].
Inter-Rater Reliability	In the context of performance measurement, inter-rater reliability assesses the percent agreement between two raters, reporting a Cohen's Kappa statistic; this method takes into consideration statistical chance. Here, one can assess the percent agreement between an automated report and manual abstraction, for instance, which is the likely scenario one would use this technique. Confidence intervals can be included with the Kappa statistic.
Test-retest Reliability	This method assesses whether responses are reproducible, at two different time points. This method can be used in instances where it is assumed that scores should be stable across time.

Face Validity	Easy to administer, straightforward, and cost- and time-efficient. Face validity assesses the degree to which measure scores can distinguish good and poor quality, among providers, at face value. NQF requires empirical tests of validity for new measures, to supplement or replace face validity testing alone.
Empirical Validity	Statistical correctness or accuracy. Sound, empirical test of validity, which is required by NQF for measure endorsement. There are several different types of empirical validity testing; convergent and discriminative validity may be used in a correlation analysis of two measures scores, to test the assumed relationship between those scores, empirically.
Correlation Analysis	If a draft measure relates to a published measure, one can assess the correlation of the two measure scores, organized by provider ID. For instance, one can test the validity of a draft measure looking at TKA readmission rates, which can be positively correlated with a published measure assessing TKA medication reconciliation rates. Here, we would expect that improvements in medication reconciliation rates should mean improvements in readmission rates, for the same or similar population.

Feasibility (Beta Testing)

Feasibility should always be assessed first, because it is not too uncommon to identify feasibility issues that will need to be addressed before other forms of testing can take place. Scientific acceptability testing requires that the performance measure is feasible, as specified. Once the measure is found to be feasible, the measure developer can move on to scientific acceptability testing.

Feasibility testing ensures[6]:

- Data availability, data accuracy, data standards, and workflow.
- The measure can be implemented without issue or undue burden.
- The measure can be captured and calculated by the EHR or registry.
- An automated report can be generated with results from the measure calculation (i.e. provider performance scores).
- Unintended consequences, and other feasibility issues (i.e. usability/intended use), are addressed.

The Table 4 below describes some of the types of feasibility testing.

Table 4. Types of Feasibility Testing.

Type of Feasibility Testing	Description of Method	Benefits of Method
Feasibility Survey & Feasibility Scorecard	Usually in survey format, the Feasibility Survey is administered to clinical staff, informatics professionals, department managers, or other health IT experts, and contains questions regarding the ability to capture all data elements and value sets, as specified by the measure. The Feasibility Survey may also contain high-level, supplementary questions regarding validity, and	• Time- and cost-efficient method for determining the availability of certain data elements and value sets in an EMR. • Both quantitative and qualitative results are gathered.

	other aspects of the measure. The Feasibility Scorecard is a scoring system, used by NQF, to summarize the results from the Feasibility Survey.	• Accepted by NQF as satisfying endorsement criteria related to feasibility.
Clinical/Provider Interviews	Interviews are typically semi-structured; a script can be followed, but flexibility should be allowed. Here, clinical staff are interviewed to gain clinical expertise related to feasibility (i.e. clinical workflow, information about how data is being captured, exclusion/exception suggestions), as well as clinical relevance.	• Can collect highly detailed and important information from clinical experts and experienced healthcare administration. • Measure Developers can be selective as to who they interview, ensuring they are collaborating with appropriate experts.
Patient/caregiver Interviews	Interviews are typically semi-structured; a script can be followed, but flexibility should be allowed. Here, patients and/or caregivers are asked questions related to the measure – its usefulness, its validity, and its general feasibility.	• Gather patient-reported information about the measure. • Can gather valuable caregiver information about the measure. • Considered by CMS and NQF as a patient-centered approach, which is supported and favored by both agencies.
Focus Groups	Structured or semi-structured group of cohorts (i.e. patients, experts, EMR vendors), who are deemed to have some special insight into the measure focus area. Focus groups can also be used in place of semi-structured interviews, in some cases.	• Excellent method for confirming information with more than one person. • Great method for including patients, to gain patient-centered insight. • Focus groups can prompt interesting discussions that otherwise would not have occurred without the group element.
BONNIE Testing	BONNIE is an online tool, developed by MITRE Corporation, a contractor of CMS, to test eCQM expression logic.	• Provides the ability to check measure logic, syntax, and readability, or one's measure logic, to ensure there are no errors. • Required by NQF for endorsement of eCQMs.

Intended Use & Unintended Consequences

In addition to feasibility, reliability, and validity, intended use and unintended consequences should be assessed during measure testing. Much of this information is gathered as part of feasibility testing. Intended use may describe which providers should report on a measure(s), or what care setting the measure should be implemented in. Unintended consequences deal with those negative consequences of measure development and implementation that were initially unseen, but significantly impact resources. Examples include: provider and reporting burden, patient burden (e.g. completing long surveys in a waiting room), decision fatigue, costly implementation (e.g. required EMR templates to be

built), or significant changes in workflow. Remember, performance measures are not truly feasible if they cause undue burden or significant unintended consequences.

Public Comment Period

Public comment of the measure(s) ensures transparency and equal input from all affected stakeholders. When a measure is released for public comment, the goal is to obtain feedback on the technical aspects of the measure, the clinical and operational implications of using the measure, and any unintended consequences. The measure, or measure set, is posted to the AAOS website for a minimum of two weeks. The public has the opportunity to review measure(s), provide comments and suggest changes. Measure developers and project managers should plan accordingly to incorporate the time needed to solicit and receive public comment, including time needed to revise the measurement set if warranted.

At the AAOS, the call for public comment on performance measures is communicated via a variety of notification outlets, including the AAOS website, AAOS publications, and sending notices to internal and external stakeholders. Requests for public comment are sent to the AAOS Board of Directors, Council on Research and Quality, Committee on Evidence-Based Quality and Value, Board of Specialty Societies, Board of Councilors, orthopaedic specialty societies, relevant medical societies, patient advocacy organizations, and health insurers. All comments are submitted to AAOS staff electronically. See Appendix D for a sample feedback request form.

After the public comment period, AAOS staff prepares a report summarizing all feedback received, including verbatim comments. The work group reviews each comment and considers measure revisions to improve clarity, while maintaining the measure focus. All public comments receive a written response. Comments and responses are included in the final measure report.

Note that the federal rulemaking process also includes a public comment period, which is distinct from the AAOS public comment period. If an AAOS-stewarded performance measure is used in a federal quality reporting program, CMS would receive feedback on that measure as well. During measure use, public comments received as part of the federal rulemaking process should be considered as part of ongoing surveillance. These comments should also be formally considered when the measure is re-evaluated.

Approval & Dissemination

Approval

If no major revisions to the measure are needed based on public comment, the measurement set goes through the process of approval. Once the measures are approved by the measure development work group, they are submitted for sequential approvals from the Committee on Evidence-Based Quality and Value, the Council on Research and Quality, and finally the Board of Directors. After Board approval, the AAOS will pursue NQF endorsement. If substantial revisions to the measure set are needed based on this approval process, the process of developing measures based on evidence review begins anew, followed by peer review, testing, soliciting public comment, and approval.

Dissemination

Work group leadership will be asked to draft a summary that will highlight the final measure(s) and rationale, as well as how providers can capture and report data for the measure. The summary will be submitted for publication in *AAOS Now* and will link to the full measure details. Following publication of the summary, AAOS staff will coordinate dissemination activities including a public notification of the release of the measures via press release, the AAOS website, social media, and other educational programs.

The primary purpose of measure development is to provide orthopaedic surgeons with tested, approved, and endorsed measures for research, outcomes documentation, and value-based payment program participation. The Measure Developer aims at having quality measures approved and implemented into quality payment programs. To ensure the utility of measures, Measure Developers must provide strong evidence to CMS that the measure(s) add value to quality reporting programs. After AAOS measures are implemented, the AAOS will monitor the performance, respond to ongoing feedback, and continuously scan the environment for similar measures.

Performance measures must be continually evaluated, as new evidence changes best practices for clinical care, new data sources become available, and/or unforeseen side effects of the measures are discovered. Sometimes the creation of a measure leads to the realization that information is not being collected or is not being collected in the best way. The AAOS will continue to maintain and modify physician level performance measures to ensure that they are consistent with the latest scientific advances. Updated information regarding physician level performance measures as well as information about the CMS QPP are continually available on the AAOS website as physician-level performance measures continue to evolve.

Measure Maintenance

Measure maintenance takes place once a measure is developed, tested, and implemented. The Measure Developer (or Steward) should follow certain schedules and methodologies, usually set forth by CMS. The types of measure maintenance are listed below. Maintenance includes NQF-endorsed and non-endorsed measures.

Annual Update

The Annual Update (AU) is a CMS-driven process where Measure Developers update their measure specifications, typically once a year, notifying CMS of any changes to their measure specifications. The primary reason for this AU process is to ensure regular updating of codes and code sets that comprise the technical specifications, although several changes can take place during AU. The measures that need to be updated are any measures that are part of CMS federal payment programs (e.g. MIPS). Retesting of measures can also take place at this time, especially if changes have been made to the measure specifications. Note that any changes to a measure's specifications warrants prompt NQF notification, if the measure is an NQF-endorsed measure. Table 5 below lists the procedures Measure Developers should keep in mind as the AU process nears[6].

Table 5. CMS-Required AU Procedures.

CMS-Required AU Procedures	Description of Method
Update Codes	Adding, deleting, updating, and accessing clinical applicability.
Assessing for Potential Harmonization/Competition	Measure Developers should assess the possibility for measure harmonization and/or competing measures. Regarding the former, justification may be needed as to why there exists competing measures.
Information Gathering	Assessing any comments or suggestions received, during the time the measure has been in use, reviewing relevant literature, ensuring no significant clinical events have occurred that can change the measure (i.e. new drugs or therapies, new research, adverse outcomes, unintended consequences). It is the responsibility of the Measure Developer to do their due diligence in the information gathering phase, to ensure their measure(s) are still relevant, safe, up-to-date, and tested.
Performance Rate Analysis	Measure Developers should assess national and regional performance rates, stratified performance rates (by procedure, gender, or whichever else is appropriate), analysis of performance gaps/variation.

CMS will evaluate the Measure Developer's AU information, and will determine the measure's disposition (e.g. Keep, Revise, Retire, Suspend). CMS will notify the Measure Developer with the next course of action. CMS also notifies NQF via report, with changes the measure may have undergone. If the measure did not undergo any updates, or if the updates are deemed negligible, the Measure Developer will not likely have to re-test their measures. Retesting of measures occurs as part of the NQF maintenance endorsement cycle, for measures that hold the endorsement designation.

NQF Continued (Maintenance) Endorsement

Measures that receive NQF endorsement maintain endorsement status for three years. At the end of three years, Measure Developers undergo the same process as do measures applying for first-time endorsement. Measures will be evaluated, alongside new measures, against NQF's Measure Evaluation Criteria.

AAOS recommends that the CMS Blueprint be closely followed, along with other NQF documents (i.e. Measure Evaluation Criteria, technical reports, schedules), as primary resources during the AU or Maintenance Endorsement process.

Conclusion

Performance measure development is a lengthy and involved process that can potentially take years to complete. Often, different stakeholders collaborate on measure development projects where there appears to be mutual and widespread interest, creating multi-organizational development groups. Here, expertise can be funneled into one measure development idea, making for highly feasible, reliable, and valid measures, but the process can become challenging and time-consuming. It is important to note that there has been an enormous amount of literature published in the past decade in performance measurement, and the AAOS encourages its members to utilize this methodology document, as well as the references supplied in Appendix E – Measure Development Resources, when developing healthcare performance measures.

AAOS
American Academy of Orthopaedic Surgeons
American Association of Orthopaedic Surgeons

References

[1] Centers for Medicare & Medicaid Services. National Health Expenditures 2016 Highlights. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/highlights.pdf. Accessed September 2018.

[2] US Department of Commerce. Bureau of Economic Analysis. https://www.bea.gov/industry/gdpbyind-data. Published 2018. Accessed September 2018.

[3] Agency for Healthcare Research and Quality. DHHS. 2011 Report to Congress: National Strategy for Quality Improvement in Health Care. https://www.ahrq.gov/workingforquality/reports/2011-annual-report.html. Accessed September 2018.

[4] Agency for Healthcare Research and Quality. DHHS. 2015 Annual Progress Report to Congress: National Strategy for Quality Improvement in Health Care. https://www.ahrq.gov/workingforquality/reports/2015-annual-report.html. Accessed September 2018.

[5] National Quality Forum. Understanding Performance Measures: Anatomy and Types. http://public.qualityforum.org/Chart%20Graphics/Understanding%20Performance%20Measures%20-%20Anatomy%20and%20Types.pdf. Published 2013. Accessed September 2018.

[6] Department of Health and Human Services. Centers for Medicare & Medicaid Services. Blueprint for the CMS Measures Management System - Version 12.0. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/Downloads/Blueprint-130.pdf. Published May 2017. Accessed September 2018.

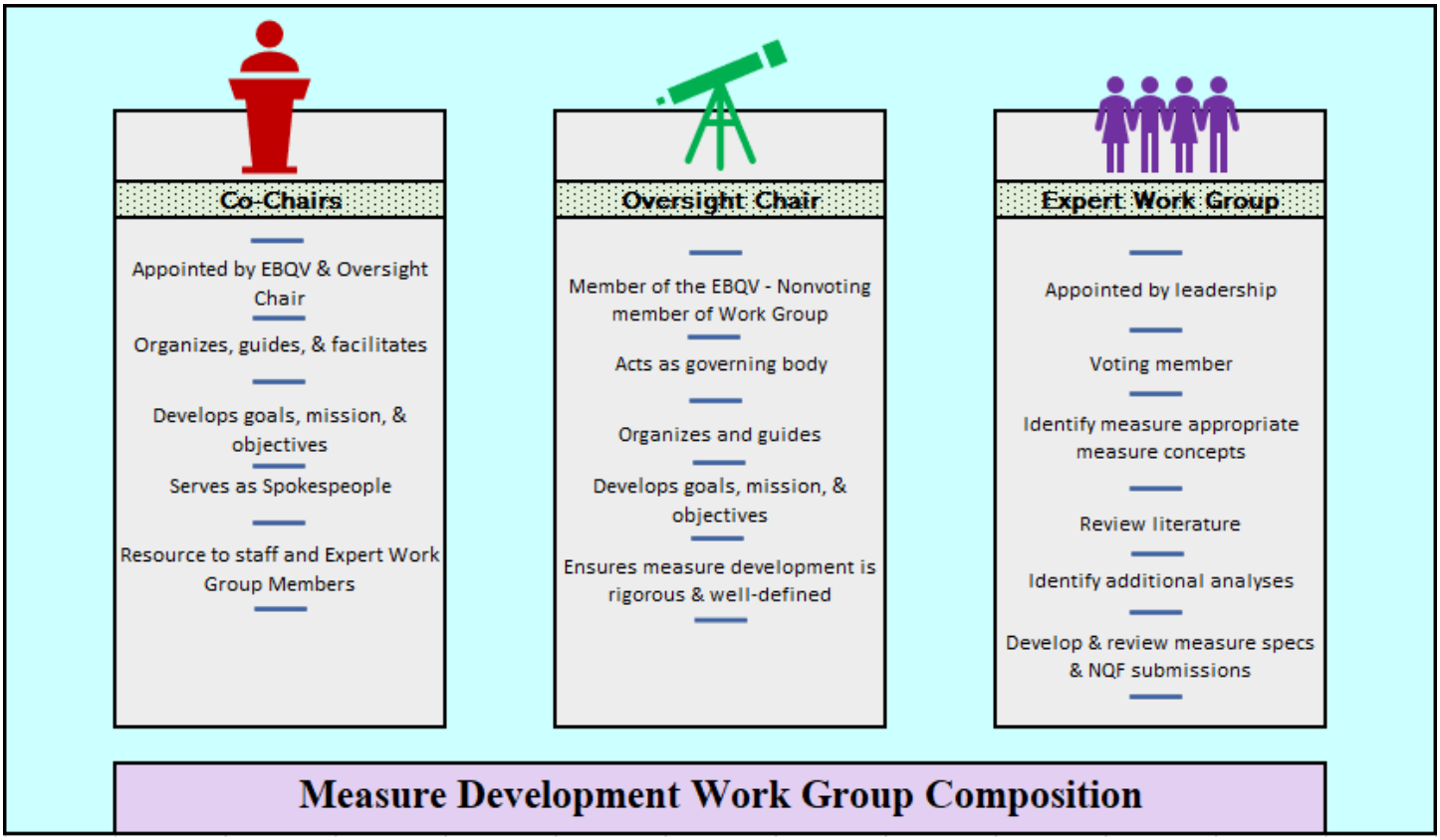
[7]. Centers for Medicare & Medicaid Services. Office of the National Coordinator for Health Information Technology. Quality Data Model, Version 4.3. https://ecqi.healthit.gov/system/files/qdm_4_3_508_compliant.pdf. Published September 27, 2016. Accessed September 2018.

[8]. Adams, J. RAND Corporation. The Reliability of Provider Profiling, A Tutorial. https://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR653.pdf. Published 2009. Accessed September 2018.

Appendix

APPENDIX A

Measure Development Work Group Composition.



APPENDIX B

Sample Criteria Matrix.

Harmonization	Importance			Feasibility	Usability
Existing Measure	Evidence	Priority (Impact)	Performance Gap	Implementation	Use
Is there an existing performance measure on this concept? If a measure currently exists, will a newly developed measure have a meaningful impact on clinical decision-making/clinical outcomes and/or reduce practice variation, and/or be of significant benefit to AAOS members?	Is there an existing evidence-based CPG or AUC on this topic? And/or are there sufficient research findings available upon which to base a clinical performance measure?	Is the topic area a national health goal/priority or high impact area of healthcare (e.g., large number of patients impacted, leading cause of morbidity and mortality, high cost, or high burden of illness)	Are there substantial variations in the diagnosis and/or treatment of the health problem?	Are there data sources for Implementation? If yes, what are they?	What are the planned uses related to this topic?

AAOS
American Academy of Orthopaedic Surgeons
American Association of Orthopaedic Surgeons

APPENDIX C

Sample Measure Specification Worksheet.

Measure #109: Osteoarthritis (OA): Function and Pain Assessment – National Quality Strategy Domain: Person and Caregiver-Centered Experience and Outcomes

2017 OPTIONS FOR INDIVIDUAL MEASURES:
CLAIMS ONLY

MEASURE TYPE:
Process

DESCRIPTION:
Percentage of patient visits for patients aged 21 years and older with a diagnosis of osteoarthritis (OA) with assessment for function and pain

INSTRUCTIONS:
This measure is to be reported at **each denominator eligible visit** occurring during the reporting period for patients with osteoarthritis seen during the reporting period. This measure may be reported by eligible clinicians who perform the quality actions described in the measure based on the services provided and the measure-specific denominator coding.

Measure Reporting:
The listed denominator criteria is used to identify the intended patient population. The numerator quality-data codes included in this specification are used to submit the quality actions allowed by the measure. All measure-specific coding should be reported on the claim(s) representing the eligible encounter.

DENOMINATOR:
All patient visits for patients aged 21 years and older with a diagnosis of OA

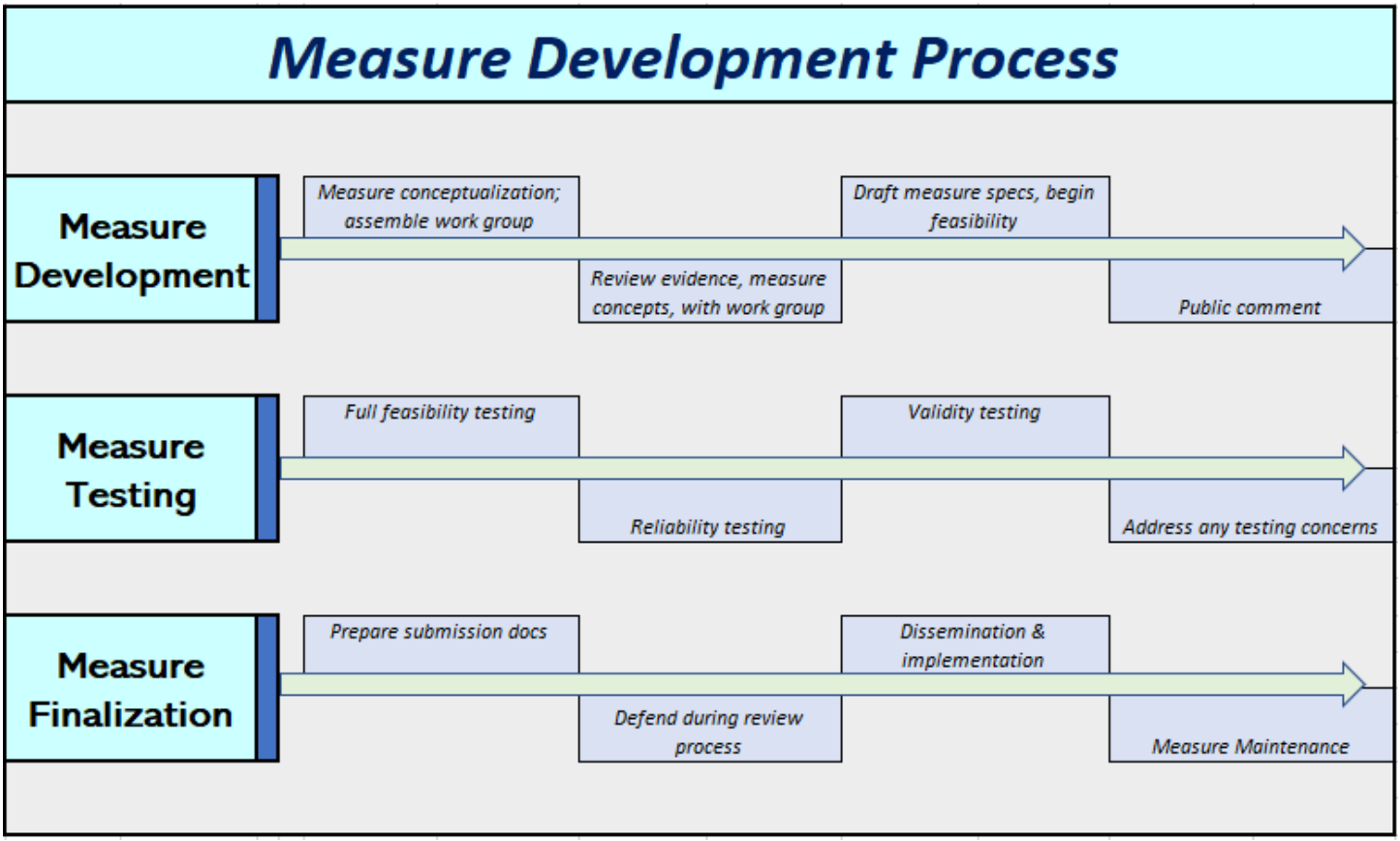
Denominator Criteria (Eligible Cases):
Patients aged ≥ 21 years on date of encounter
AND
Diagnosis for osteoarthritis (OA) (ICD-10-CM): M15.0, M15.1, M15.2, M15.3, M15.4, M15.8, M15.9, M16.0, M16.10, M16.11, M16.12, M16.2, M16.30, M16.31, M16.32, M16.4, M16.50, M16.51, M16.52, M16.6, M16.7, M16.9, M17.0, M17.10, M17.11, M17.12, M17.2, M17.30, M17.31, M17.32, M17.4, M17.5, M17.9, M18.0, M18.10, M18.11, M18.12, M18.2, M18.30, M18.31, M18.32, M18.4, M18.50, M18.51, M18.52, M18.9, M19.011, M19.012, M19.019, M19.021, M19.022, M19.029, M19.031, M19.032, M19.039, M19.041, M19.042, M19.049, M19.071, M19.072, M19.079, M19.111, M19.112, M19.119, M19.121, M19.122, M19.129, M19.131, M19.132, M19.139, M19.141, M19.142, M19.149, M19.171, M19.172, M19.179, M19.211, M19.212, M19.219, M19.221, M19.222, M19.229, M19.231, M19.232, M19.239, M19.241, M19.242, M19.249, M19.271, M19.272, M19.279, M19.90, M19.91, M19.92, M19.93
AND
Patient encounter during the reporting period (CPT): 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215

NUMERATOR:
Patient visits with assessment for level of function and pain documented (may include the use of a standardized scale or the completion of an assessment questionnaire, such as an SF-36, AAOS Hip & Knee Questionnaire)

NUMERATOR NOTE: *For the purposes of this measure, the method for assessing function and pain is left up to the discretion of the individual eligible clinician and based on the needs of the patient. The assessment may be done via a validated instrument (though one is not required) that measures pain and various functional elements including a patient's ability to perform activities of daily living (ADLs).*

Appendix D

Measure Development Process.



APPENDIX E

Measure Development Resources.

General Resources
CMS Blueprint, V13.0, May 2017 [general]
CMS eHealth Homepage
NIH: Value Set Authority Center (VSAC) [identifying value sets, measure retooling]
Office of the National Coordinator for Health Information Technology (ONC) [HIT]
CMS – Quality Payment Program home page (QPP)
CMS/QPP Final Rule Executive Summary/Proposed Y3 [information for MACRA/MIPS/QPP] QPP Y2 QPP Y3

National Quality Forum (NQF) – Homepage [measure endorser]
CMS and NQF Measure Inventory Tools [look up measures] CMS NQF
eCQMs
eCQM Resource Center [eMeasure information]
ONC eCQM 101
CMS Measure Authoring Tool (MAT) Information Page [specifying measures]
CMS/JIRA [tracking/submitting tickets, following PM technologies]
Bonnie/MIRTE [test eCQM logic]
Clinical Quality Language (CQL)
CQL Github [new healthcare expression language]
CQL Github [resource #2]
Health Level Seven International (HL7) QCL Information Page
eCQM Resource Center for CQL
CMS CQL Basics – YouTube
Measure Development [Specs and Testing]
Adams Reliability Paper
NQF Technical Papers
eCQM Measure Development Resources
CMS webinar, MAT User Information