A Blueprint for the CMS Measures Management System

Version 9.1

Volume 1 of 2

December 2012

Prepared by: Health Services Advisory Group, Inc.
3133 E Camelback Road, Ste 300
Phoenix, AZ 85016

Phone: 602.264.6382
Fax: 602.241.0757
Web: www.hsag.com

Table of Contents

1. Introduction 1-1

1.1 Background 1-1

1.2 Development and Maintenance of the CMS Measures Management System Blueprint 1-2

1.3 Measures Management System Framework 1-4

1.4 Why Version 9? 1-5

1.5 Structure of the Blueprint 1-6

1.6 Role of the Measure Contractor 1-7

1.7 Role of the Measure Contractor's COR/GTL 1-8

1.8 Role of the Measures Manager 1-9

2. Measure Priorities Planning 2-1

2.1 Introduction 2-1

2.2 Inputs into CMS Priorities Planning 2-2

2.3 Quality Measure Development and Endorsement Projects 2-6

2.4 Pre-Rulemaking Process 2-6

2.5 CMS Considers MAP Input for Final Selection 2-8

2.6 CMS Rulemaking Processes 2-8

2.7 Rollout, Production, and Monitoring of Measures 2-9

2.8 Measures Maintenance and Impact Assessment 2-9

2.9 How the Measure Manager can assist CMS 2-10

2.10 How the Measures Manager can assist the Measure Contractors 2-10

3. Measure Development 3-1

3.1 Introduction 3-1

3.2 Deliverables 3-4

3.3 Types of Measures 3-4

3.4. Procedure 3-5

4. Measure Evaluation 4-1

4.1 Introduction 4-1

4.2 Deliverables 4-2

4.3 Discussion of Measure Evaluation Criteria and Subcriteria 4-2

4.4 Applying the Measure Evaluation Criteria 4-6

4.5 Using the Measure Evaluation Report and Measure Evaluation Guidance 4-8

4.6 Procedure 4-8

4.7 Special Considerations 4-9

4.8 Evaluation of Measure Set 4-13

4.9 Overview of Appendices 4-14

5. Harmonization 5-1

5.1 Introduction 5-1

5.2 Procedure 5-2

6. Information Gathering 6-1

6.1 Introduction 6-1

6.2 Deliverables 6-2

6.3 Procedure 6-2

7. Technical Expert Panel 7-1

7.1 Introduction 7-1

7.2 Deliverables 7-2

7.3 Procedure 7-2

8. Technical Specifications 8-1

8.1 Introduction 8-1

8.2 Deliverables 8-3

8.3 Procedure 8-3

8.4 Special Considerations 8-16

9. eMeasure Specifications 9-1

9.1 Introduction 9-1

9.2 Deliverables 9-3

9.3 Health Quality Measures Format 9-3

9.4 National Quality Forum Quality Data Model 9-4

9.5 Building Block Approach to eMeasures 9-5

9.6 Measure Authoring Tool 9-5

9.7 Procedure 9-5

9.8 Preparing eMeasure Specifications for Public Release 9-32

9.9 eMeasure Future and Next Steps 9-34

10. Special Topics 10-1

10.1 Introduction 10-1

10.2 Cost and Resource Use Measures 10-2

10.3 Composite Measures 10-7

10.4 Outcome Measures 10-12

10.5 Multiple Chronic Conditions Measures 10-16

11. Risk Adjustment 11-1

11.1. Introduction 11-1

11.2 Deliverables 11-2

11.3 Attributes of Risk Adjustment Models 11-2

11.4 Procedure 11-5

12. Public Comment 12-1

12.1 Introduction 12-1

12.2 Deliverables 12-1

12.3 Procedure 12-1

13. Measure Testing 13-1

13.1 Introduction 13-1

13.2 Deliverables 13-1

13.3 Alpha and Beta Testing 13-1

13.4 Procedure 13-4

13.5 Testing and Measure Evaluation Criteria 13-10

13.6 Special Considerations 13-18

14. National Quality Forum Endorsement 14-1

14.1 Introduction 14-1

14.2 Procedure 14-1

14.3 NQF 2-Stage Consensus Development Process (CDP) 14-9

14.4 Measure Maintenance 14-10

15. Measure Rollout 15-1

15.1 Introduction 15-1

15.2 Deliverables 15-1

15.3 Procedure 15-2

16. Glossary 16-1

HSAG HEALTH SERVICES ADVISORY GROUP

Blueprint Version 9.0

In response to the rapidly changing environment surrounding health care quality measures, Health Services Advisory Group (HSAG), the CMS Measures Manager, has updated the *Blueprint for the CMS Measures Management System, Version 8.0* (the Blueprint). Version 9 of the blueprint is divided into two volumes, one for measure development and the for measure maintenance.

Section in Version 9.0, Volume 1, Measure Development	Description of Change
Section 1: Introduction	Added/Changed: • Streamlined the "Background" section. • Additional information about the process of updating the Blueprint. • Added "Significant Changes in Version 9" section.
Section 2: Measure Priorities Planning	Added/Changed: • The majority of the section was revised. • More focus on measure alignment and harmonization across CMS and HHS. • Added an explanation of the measure contractors' roles in measure planning and prioritization.
Section 3: Harmonization	New section within the Blueprint, information previously contained in other sections but consolidated due to importance of the topic.
Section 4: Measure Development	Added/Changed: • Updated many of the flowcharts. • Added information regarding deliverables for eMeasures. • Updated outcome measure definition. • Provided further clarification regarding the kick off meeting with the Measures Manager. • Updated the information about how the TEP evaluates or reviews measures. • Provided information regarding how measures submission has changed from Version 8.0.
Section 5: Measure Evaluation	Added/Changed: • Harmonization included as 5th measure evaluation criteria for CMS contractors. • Usability criterion was updated to "Usability and Use".

Section in Version 9.0, Volume 1, Measure Development	Description of Change
	• Guidance and tools for evaluating measures for Usability and Use (criteria and sub-criteria) were updated to align with NQF updated criteria. • Added guiding principles that can be used to evaluate a measure set.
Section 6: Information Gathering	Added/Changed: • Further guidance on the process of information gathering. • Guidance related to the Paper Reduction Act. • The number of Steps in the information gathering process. For better read and organization, collapsed a few steps into a single step. • Relocated FAQ responses to the body of the section.
Section 7: Technical Expert Panel	Added/Changed: • Guidance on TEP for other type of contracts besides measure development contract. • Further guidance on the process of evaluating the measures . • Guidance on public attendance at the TEP. • Information about posting process for Call for TEP. • Included language regarding the Paperwork Reduction Act (PRA).
Section 8: Technical Specifications	Added/Changed: • Made changes to language referencing MAT development for eMeasures. • Aligned the section with definitions in the eMeasure Specifications section. • Reordered Exclusions/Exceptions discussion to promote clarity. • Clarified measure data sources with respect to EHRs. • Added information about use of ambiguous language in time references during the development of measures. • Updated NQF ICD-10- CM/PCS timeline related to measure submission as a result of CMS' proposed delay in implementation. • Harmonization was moved to a standalone section within the Blueprint. • FAQ- relocated to body of the section.
Section 9: eMeasure Specifications	Added: • Information on development process • Description- both verbal and visual- of the icon located throughout the Blueprint denoting special considerations for eMeasures. • Descriptive (verbal and visual) of the technical specifications development process and factors influencing the development. • Added a list of Deliverables due to the COR/GTL at the completion

Section in Version 9.0, Volume 1, Measure Development	Description of Change
	of eMeasures Specifications development. • Metadata table (Table 8-1)- added field for eMeasure Identifier (Measure Authoring Tool). • Added brief statement regarding ambiguities in time references within eMeasures. • Added information regarding NLM value set management process. • Added Appendix 8-C with detailed information on performance calculation of eMeasures. • Information regarding "delta" measures. Updated • Modified the Figure (now 8-2) depicting the process of eMeasure Specifications development and transmission format and added steps for determining list of measures, obtaining final COR/GTL approval, and submitted the eMeasure for NQF endorsement. • Updated the ONC HIT Standards Committee vocabulary standards- and provided 2 tables from different starting points. • Updated process for requesting new SNOMED CT concepts. • Additional information on exclusions and exceptions in eMeasures. • Value sets for supplemental data elements. • Sample eMeasure to reflect changes in metadata fields, and correctly reflect QDM references. • Acronym and abbreviations list.
Section 10: Special Topics	New section in the Blueprint that combines prior sections (Outcomes, Composite, and Cost and Resource Use measures), and now includes a discussion on Multiple Chronic Conditions.
Section 11: Risk Adjustment	Added/Changed: • Significant changes to Table 11.1 to improve the descriptions of the attributes of risk adjustment models. • Added more detail to sample definition. • Added information regarding complex or new statistical models .throughout the section. • Enhanced description of the risk adjustment methodology report. • Added citations and reference material throughout the section. • Minor editorial changes throughout the section to improve clarity.
Section 12: Public Comment	Added/Changed: • Revised the MMS posting process to reflect the changes outlined by CMS.
Section 13: Measure Testing	Added/Changed: • Measure testing plan as a deliverable.

Section in Version 9.0, Volume 1, Measure Development	Description of Change
	• Table describing key features of alpha and beta testing. • Additional information on the training and qualifications of measure developer staff. • Minor updates to clarify the reliability and validity sections, including updated references. • Symbols denoting information relevant to eMeasures provided in Blueprint section. • Information based on NQF Draft Requirements for eMeasure Testing. • Additional guidance provided for feasibility testing for eMeasures.
Section 14: National Quality Forum Endorsement	Added/Changed: • Added information about the proposed NQF 2-Stage CDP process. • General changes to promote clarity.
Section 15: Measure Rollout	Added/Changed: • Changed Figure 15-1. • Added some language about the MAP and the pre-rulemaking process. • Updated language regarding the Paperwork Reduction Act. • Added information regarding the auditing and validation plan and appeals process. • Supplemented the business processes step with information regarding VBP and P4R.
Section 16: Glossary	Additional terms reflective of the changes throughout version 9, volume 1 of the Blueprint.

1. Introduction

1.1 Background

The Centers for Medicare & Medicaid Services (CMS) has developed a standardized approach for developing and maintaining the quality measures used in its various quality initiatives and programs. Known as the Measures Management System, this system is composed of a set of business processes and decision criteria that CMS funded measure developers (or contractors) follow when developing, implementing and maintaining quality measures. The major goal of the Measures Management System is to provide sufficient information to the measure developers to help them produce high caliber quality measures that are appropriate for accountability purposes. The Measures Management System was developed to help CMS manage an ever increasing demand for quality measures to use in its various public reporting and quality programs as well as in value-based purchasing initiatives. The full Measures Management System set of business processes and decision criteria are documented in or described in this manual, the CMS Measures Management System Blueprint, Version 9.0 (the Blueprint).

When issuing contracts for measure development and maintenance, CMS must show how the recommended measures will relate to the goals and objectives set forth by various national action plans and strategies such as the *National Strategy for Quality Improvement in Health Care* (the National Quality Strategy), the *National Prevention and Health Promotion Strategy*, and the *National Action Plan to Prevent Healthcare-Association Infections: Roadmap to Elimination*.

The Secretary of the Department of Health and Human Services (HHS) submitted the *Annual Progress Report to Congress on the National Quality Strategy* in March, 2012. This report updates the initial (2011) National Quality Strategy that established three aims and six priorities for quality improvement.[1] The six priorities listed in the annual report are:

- Making care safer by reducing harm caused in the delivery of care.
- Ensuring that each person and his or her family members are engaged as partners in their care.
- Promoting effective communication and coordination of care.
- Promoting the most effective prevention and treatment practices for the leading causes of mortality, starting with cardiovascular disease.
- Working with communities to promote wide use of best practices to enable healthy living.
- Making quality care more affordable for individuals, families, employers, and governments by developing and spreading new health care delivery models.

CMS supports these goals by gathering data about the quality of care provided to Medicare beneficiaries, aggregating that information, and reporting feedback to health care providers and others.

[1] *National Strategy for Quality Improvement in Health Care*. Available at: http://www.ahrq.gov/workingforquality/nqs/nqs2012annlrpt.pdf. Accessed on July 12, 2012

CMS has long played a leadership role in quality measurement and public reporting. CMS started by measuring quality in hospitals and dialysis facilities, and now measures and publicly reports the quality of care in nursing homes, home health agencies, and physician offices. Beginning in 2012, CMS efforts will expand the quality reporting programs to include inpatient rehabilitation facilities, inpatient psychiatric facilities, cancer hospitals, and hospice programs. CMS is also transforming from a passive payer to an active value purchaser by implementing payment mechanisms that reward providers who achieve better quality or improve the quality of care they provide.

Measures Manager

CMS contracts with external organizations to assist in the development and implementation of quality measurement programs. These include Quality Improvement Organizations (QIOs), university groups, heath services research organizations, and consulting groups. CMS also contracts with the Health Services Advisory Group (HSAG) to function as the Measures Management Team (Measures Manager). The Measures Manager assists the CMS Contracting Officer Representatives/Government Task Leaders (COR/GTL) and their various contractors in their work implementing the Measures Management System. The Measures Manager role will be delineated in further detail at the end of this section.

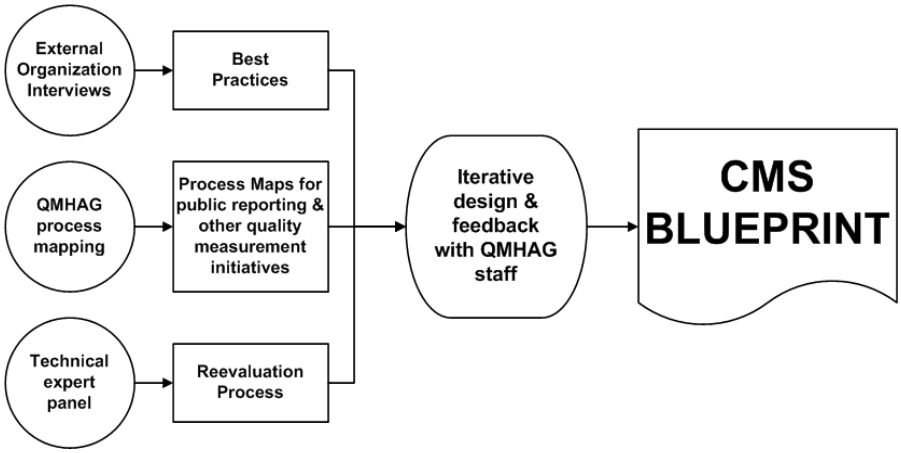
1.2 Development and Maintenance of the CMS Measures Management System Blueprint

Original design

CMS launched a project in October 2003 to design and implement the Measures Management System. As shown in Figure 1-1, the CMS Measures Management System Blueprint was designed based on the quality measurement work at CMS, augmented by the best practices of other major measure developers. Sound business process management principles, as exemplified by the Malcolm Baldrige Award criteria and Lean methodology, were also incorporated.

Structured interviews were conducted with CMS staff as well as major measure developers, and a series of process maps were developed. Simultaneously, a technical expert panel (TEP), consisting of representatives from the major measure developers, quality measurement experts, and major purchasing alliances, was convened to assist in developing the framework for a reevaluation process, including the frequency, depth, and parameters of the review.

Figure 1-1 Development of the CMS Measures Management System Blueprint

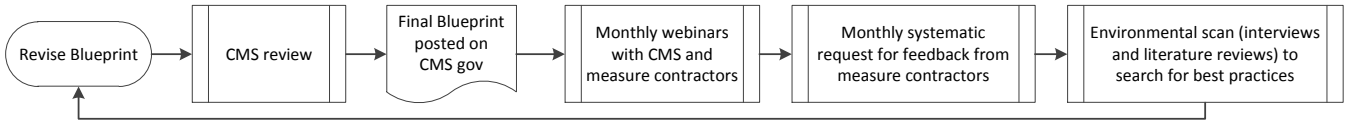


The best practices of other major measure developers were gathered through multiple telephone interviews conducted with the key personnel of these organizations using a set of standardized interview questions. Flow charts were developed based on the information obtained, which were later sent back to the organization personnel for review. Additional telephone interviews were conducted as needed. Each organization reported its own best practices or strengths.

Updates to the Blueprint

The Blueprint is updated annually by the Measures Manager. Updates are necessary because of the evolving nature of the quality measurement environment. Several pieces of legislation have affected CMS's use of quality measures. Updates to the Blueprint have incorporated the effects of these laws. In addition to paying attention to new processes incorporated by major measure organizations, the Measures Manager systematically solicits feedback and suggestions from the end users of the Blueprint. Figure 1-2 explains the processes used to update the Blueprint.

Figure 1-2 Updating the Blueprint



Structural changes (providing separate volumes for measure development and maintenance, streamlining forms, and ensuring alignment with the National Quality Forum processes and measure submission form) have been made to the Blueprint to improve its usability.

In developing and maintaining the Measures Management System, the Measures Manager uses guiding principles. These principles include making the decision-making processes involved with developing and maintaining measures transparent and ensuring that input is sought from stakeholders at multiple points along the measure development and maintenance life cycle. There is also a need for clear accountability and the roles and responsibilities of CMS, the measure contractors, and the Measures Manager must be understood and recognized. The Blueprint standardizes the processes

involved with measure development and maintenance. Communication and collaboration is increasingly important between CMS and its measure contractors, as well across other HHS agencies and with external measure developers, as the quality measurement environment increases in scope and complexity.

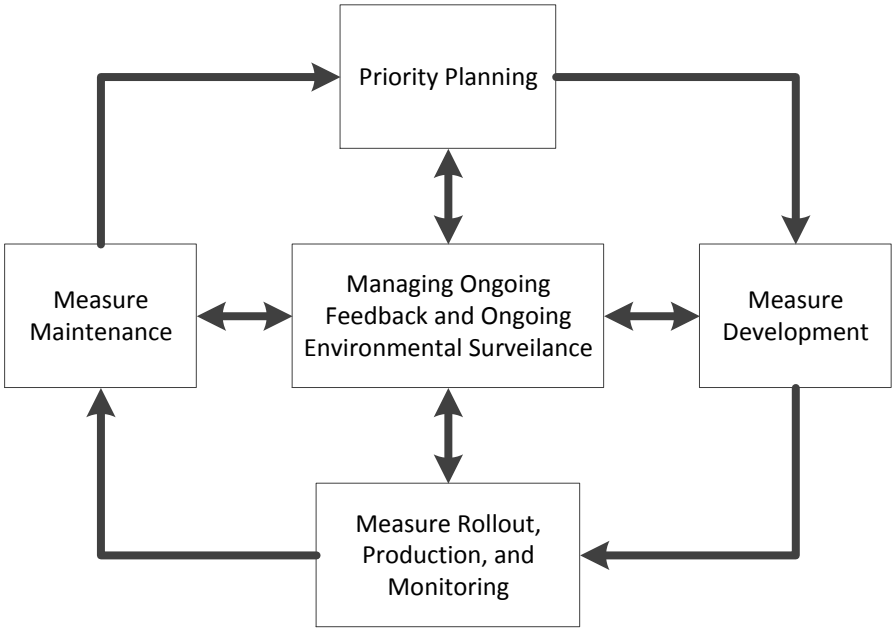
Two significant legislative Acts have recently affected quality measurement and led to Blueprint updates.

- The Patient Protection and Affordable Care Act of 2010 (ACA) provides additional funding for the creation of a wide array of quality measures, including outcome measures and measures for settings that are new to quality reporting, such as the Inpatient Rehabilitation Facilities, Hospices, Long Term Care Hospitals, Psychiatric units and hospitals, and Cancer hospitals. In addition, new quality measurement activities are being implemented for Medicaid and other HHS programs.
- The American Recovery and Reinvestment Act of 2009 (ARRA) provides significant funding for the development of standards for electronic health records (EHR) and the widespread adoption and meaningful use of EHR systems across providers. This is an area of measure development that is gaining momentum and it is now feasible to collect and report measures from EHRs.

1.3 Measures Management System Framework

The CMS Measures Management System model is shown in Figure 1-3. It is a framework to comprehensively evaluate the key processes in the measure lifecycle.

Figure 1-3 CMS Measures Management System Model



This Measures Management System framework proposes that management of measures includes the following major categories of activities: priority planning, development, rollout, production and

monitoring, and maintenance. These activities are augmented by another component, which are the ongoing surveillance of the environment and the ongoing management of feedback on the measures or measure sets. Each of these measure development, implementation, and maintenance activities contain specific tasks that are performed either sequentially or concurrently.

1.4 Why Version 9?

As stated earlier, there have been significant changes in the quality measurement environment, both in the way measures are developed, maintained, and endorsed, as well as those changes mandated by legislation.

Significant changes in Version 9

The following list summarizes significant changes made in Version 9.0. Other, more granular changes that were made in each of the individual sections are noted in the "Changes Made in Blueprint, Version 9.0" document, which is located after the Table of Contents.

- Revised the measure priorities planning section to outline CMS efforts to align and harmonize across care settings as well across programs implemented by HHS. Information has also been added to this section about the pre-rulemaking processes required by the ACA.
- Updated the requirements of NQF's consensus development process (CDP), including information about the pilot 2-stage process, the latest iteration of the NQF Measure Submission Form, and the updated NQF measure maintenance policies.
- Simplified the Measure Information Form (MIF) and Measure Justification to align with NQF's online measure submission process.
- Updated measure evaluation criteria to align with NQF criteria.
- Enhanced guidance on the development of eMeasures, including the use of an icon denoting special considerations for eMeasures during development and maintenance of measures.
- Added a new section on special topics that combines prior sections (Outcomes, Composite, and Cost and Resource Use measures), and now includes a discussion on Multiple Chronic Conditions.
- Clarified instructions for using the CMS Measures Management System Web site to issue calls for measures, Technical Expert Panels, and public comment.
- Added a separate section to highlight the need for measure alignment and harmonization.
- Enhanced the Information Gathering section with added focus on the National Quality Strategy priorities and other quality goals.
- Enhanced information regarding the Paperwork Reduction Act in relevant sections.
- Enhanced discussion on the role of ongoing measure testing during the maintenance of implemented measures.

1.5 Structure of the Blueprint

The Blueprint is divided into two volumes. The first volume, Measure Development, documents the various processes necessary to plan, develop, test, and roll out a measure. The second volume,

Measure Maintenance, documents the processes for production, monitoring, and maintaining a measure over time. The specific sections of each volume are listed below.

Volume 1: Measure Development

Section 1.	Introduction
Section 2.	Measure Priorities Planning
Section 3.	Measure Development
Section 4.	Measure Evaluation
Section 5.	Harmonization
Section 6.	Information Gathering
Section 7.	Technical Expert Panel
Section 8.	Technical Specifications
Section 9.	eMeasure Specifications
Section 10.	Special Topics
Section 11.	Risk Adjustment
Section 12.	Public Comment
Section 13.	Measure Testing
Section 14.	National Quality Forum Endorsement
Section 15.	Measure Rollout
Section 16.	Glossary

Volume 2: Measure Maintenance

Section 1.	Introduction
Section 2.	Measure Evaluation During Maintenance Phase
Section 3.	Harmonization During Maintenance Phase
Section 4.	Measure Production and Monitoring
Section 5.	Measure Maintenance
Section 6.	Measure Update
Section 7.	Comprehensive Reevaluation
Section 8.	Ad Hoc Review
Section 9.	Information Gathering During Maintenance Phase
Section 10.	Technical Specification During Maintenance Phase
Section 11.	eMeasure Specifications During Maintenance Phase
Section 12.	Technical Expert Panels During Maintenance Phase
Section 13.	Public Comment During Maintenance Phase
Section 14.	Measure Testing During Maintenance Phase
Section 15.	National Quality Forum Endorsement Maintenance
Section 16.	Glossary

1.6 Role of the Measure Contractor

The measure contractor is responsible for the development, implementation, and maintenance of the measures, as required by his or her contract with CMS. The CMS-approved processes are described in the Measures Management System Blueprint. Tools to assist and guide the measure contractor are provided with each section. These tools are intended to be integrated into their work and to produce materials that serve as deliverables to inform CMS and document contractors' progress.

The measure contractor, in developing and/or maintaining measures, will:

- Use the processes and forms shown in the Measures Management System Blueprint.
- Consult with the Measures Manager as needed to explain the use of the Blueprint.
- Assess the Blueprint in the context of the measure contract and good business practice. If the Blueprint appears to contradict the contract or appears to require additional work with minimal or no additional value, the measure contractor shall discuss the situation with his or her COR/GTL, and/or the Measures Manager.
- eMeasures must conform to the HL7 Health Quality Measures Format and have correctly mapped data criteria to the NQF Quality Data Model.
- Ensure that all relevant deliverables are provided to the COR/GTL and comply with Section 508 Amendment to the Rehabilitation Act of 1973.[2]
- Provide feedback on the use of the Blueprint to his or her COR/GTL and the Measures Manager.

1.7 Role of the Measure Contractor's COR/GTL

Within the context of the Measures Management System, the measure contractor's COR/GTL is ultimately responsible for the successful completion of the tasks in the measure development and maintenance contracts. Specifically, this includes:

- Understanding the Measures Management System Blueprint, and how it relates to the work of the measure contractor.
- Ensuring that the relevant sections of the Measures Management System Blueprint and required deliverables are incorporated into the request for proposals (RFP), task orders, or other contracting vehicles and the ensuing contract appropriately.
- Notifying the Measures Manager when a new measure contract is awarded.
- Supporting basic training and providing first-line technical assistance to the measure contractor for the Measures Management System Blueprint.
- Requiring the measure contractor's compliance with the Measures Management System Blueprint when appropriate.
- Determining when deviation from the Measures Management System is appropriate and providing or obtaining CMS authorization for this deviation. This may be done in consultation with the Measures Manager and/or the Measures Manager's COR/GTL as well as CMS management.
- Providing or obtaining CMS approval of the measure contractor's work at the specified points in the Measures Management System Blueprint.
- Contacting the Measures Manager and/or the Measures Manager's COR/GTL with any questions about the Measures Management System Blueprint, or directing the measure contractor to do so.
- Providing updates to the Measure Manager for the CMS Measure Inventory. This includes the measures concepts considered for development, measures being developed, any measures no longer being considered or developed, and information regarding NQF submission.

[2] http://www.hhs.gov/web/508/index.html

- Providing feedback on the use of the Blueprint.
- Notifying the Measures Manager GTL when a contract has ended.

A companion document, The GTL Handbook, has been developed to help COR/GTLs understand their role.

1.8 Role of the Measures Manager

The Measures Manager assists CMS and its measure contractors as they use the Measures Management System Blueprint to develop, implement, and maintain the health care quality measures. The Measures Manager fulfills this mission by:

- Supporting CMS in its work of prioritizing and planning measure activities and quality initiatives.
- Offering technical assistance to measure contractors and CMS during measure development and monitoring processes.
- Providing expertise and a crosscutting perspective to CMS and measure contractors regarding measures and measurement methods and strategies.
- Continually scanning the measurement environment to ensure CMS is informed of issues related to the quality measures in a timely fashion.
- Leading efforts to identify opportunities for harmonization of measures and measure activities across settings of care.
- Serving as an unbiased focal point to facilitate harmonization of different measure sets or to assist in evaluation of different measures for inclusion in a program or initiative.
- Assisting CMS in its liaison and harmonization work with multiple internal CMS and external key organizations: NQF, quality alliances, major measure developers and AHRQ. This assistance is critical in establishing consensus on measurement policies, coordinating measure inventories, and promoting measure harmonization.
- Informing CMS of new developments in the quality measurement environment, thereby enabling them to continue to be an effective leader in improving quality of care.
- Soliciting feedback from the measure contractors and CMS as to the success of the Measures Manager in fulfilling its mission and seeking input in new areas where the Measures Manager can provide support.
- Conducting continuous refinement of the Measures Management System based on the evolving needs of CMS, customer feedback, and ongoing changes in the science of quality measurement.
- Conducting informational sessions on updates to the Blueprint.
- Ensuring that the Blueprint and related Web-based deliverables comply with Section 508.
- Ensuring, that to the extent possible, that the Measures Management System processes are aligned with NQF requirements.

2. Measure Priorities Planning

2.1 Introduction

The Centers for Medicare & Medicaid Services (CMS) conduct measures priorities planning to establish the Medicare program's quality measurement agenda for the next five to ten years. CMS is responsible for managing activities such as measure development, testing, maintenance, implementation, and public reporting. These activities cover the spectrum of health care service delivery settings such as hospitals, outpatient facilities, physician offices, nursing homes, home health agencies, hospices, inpatient rehabilitation facilities, and dialysis facilities.

In broad terms, CMS frames its measure development and maintenance work on the three National Quality Strategy aims of better quality of care, better health, and lower costs and the six priorities: patient safety, care coordination, population and community health, efficiency and cost reduction, person- and caregiver- centered experience and outcomes, and clinical quality of care. These focus areas drive measure development, selection, and implementation activities.

Alignment, harmonization, and prioritization

CMS is working to align, harmonize, and prioritize measure development and maintenance activities across programs and settings. CMS convenes a number of working groups to look at measures and measure sets to determine where opportunities exist for alignment and harmonization. CMS is also actively engaged with groups at the Department of Health and Human Services (HHS) level and with the National Quality Forum (NQF) convened Measure Applications Partnership (MAP). CMS' goal is develop core measure sets and, to the extent possible, align these core measure sets across providers, programs, and settings.

Measure contractor

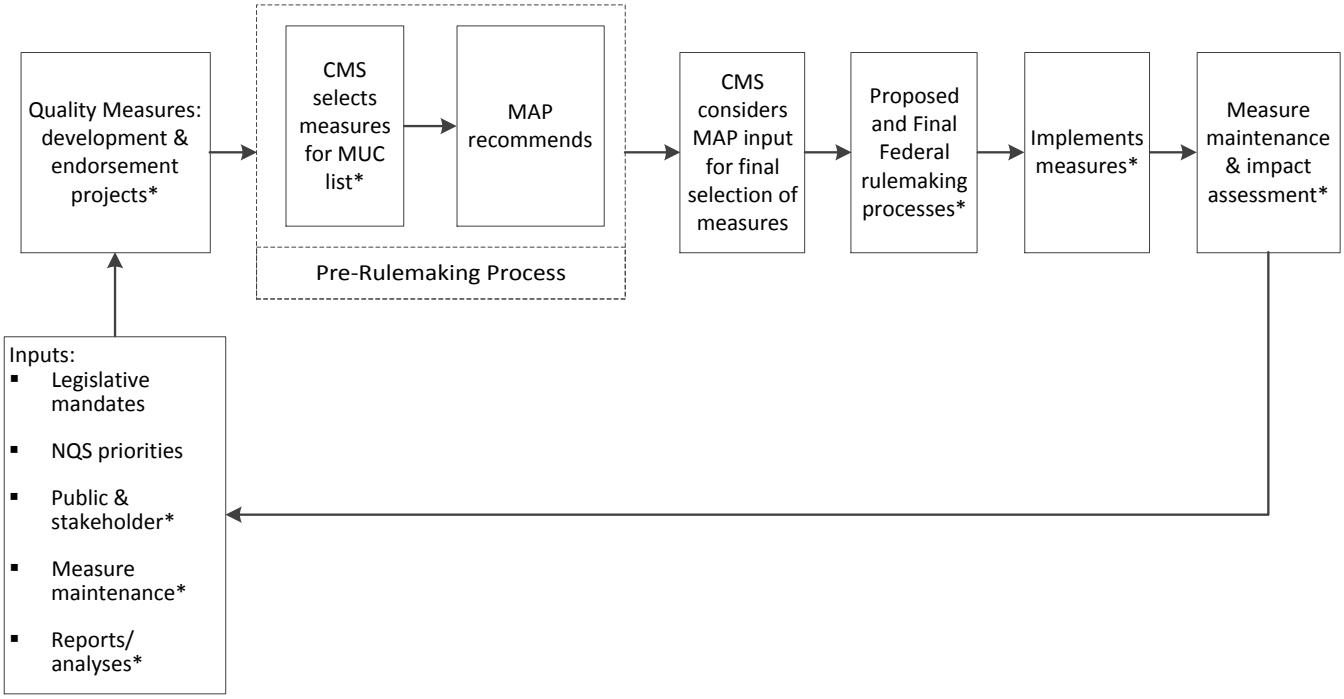
While the majority of the work outlined in this section explains CMS's role in planning and evaluating measurement activities, there are numerous places where the CMS measure contractors may undertake key activities. It is important for measure contractors to be knowledgeable about how CMS plans its measurement development and maintenance activities so that measure harmonization and alignment is achieved to the greatest degree possible.

During measure development, it is important that the measure contractors conduct a thorough environmental scan and are knowledgeable about measures that may be related or similar to those they are contracted to develop. To the extent possible, contractors are to avoid developing competing measures—those that essentially address the same concepts for the target process, condition, event or outcome and the same target patient population. Competing measures are the same at the conceptual level, but may differ in technical specifications.

During measure maintenance, it is important that the measure contractors analyze the measure performance trends, determine if this is still the best or most relevant measure, and determine if there are unintended consequences that need to be addressed.

Figure 2-1 outlines the CMS Quality Measure Selection and Implementation Activities. Boxes marked by an asterisk are those processes and activities where measure development and maintenance contractors can be involved and where they play key roles. Following the graph are descriptions of the various processes, activities, and workgroups that are involved in this complex process.

Figure 2-1 Quality Measure Selection and Implementation



2.2 Inputs into CMS Priorities Planning

Legislative mandates

Legislation can also provide input into CMS measures priority planning. Two important and recent acts that have had a strong influence on CMS's priority planning include:

- The Patient Protection and Affordable Care Act of 2010 (ACA) provides additional funding for the creation of a wide array of quality measures, including outcome measures and measures for settings that are new to quality reporting, such as the Inpatient Rehabilitation Facilities, Hospices, Long Term Care Hospitals, Psychiatric units and hospitals, and PPS-Exempt Cancer hospitals. In addition, new quality measurement activities are being implemented for Medicaid and other HHS programs.
- The American Recovery and Reinvestment Act of 2009 (ARRA) provides significant funding for the development of standards for electronic health records (EHRs) and the widespread adoption and meaningful use of EHR systems across providers. This is an area of measure development that is gaining momentum and it is now feasible to collect and report measures from EHRs.

Measure contractor

Measure contractors should be knowledgeable of legislative mandates, particularly those impacting their measures and/or scopes of work.

National Quality Strategy

The National Strategy for Quality Improvement in Health Care (the National Quality Strategy) sets a course for improving the quality of health and health care for all Americans. It serves as a blue print for health care stakeholders across the country – patients, providers, employers, health insurance companies, academic researchers, and local, State, and Federal governments – that helps prioritize quality improvement efforts, share lessons, and measure our collective success.

The initial National Quality Strategy, published in March 2011, established three aims and six priorities for quality improvement. This report was updated in March 2012[1] and details some of the work conducted in public and private sectors over the past year to advance and further refine those aims and priorities. This report also focuses attention on the aims and priorities by including key measures that HHS will use to evaluate the Nation's progress towards the quality improvement aims of the National Quality Strategy. The report also provides concrete examples of new initiatives at HHS, and among other public and private stakeholders, that are directly working to advance the Strategy's goals.

National Quality Strategy's three aims

1. Better Care—Improve the overall quality of care, by making health care more patient-centered, reliable, accessible, and safe.
2. Healthy People/Healthy Communities—Improve the health of the U.S. population by supporting proven interventions to address behavioral, social, and environmental determinants of health in addition to delivering higher-quality care.
3. Affordable Care—Reduce the cost of quality health care for individuals, families, employers, and government.

National Quality Strategy's six priorities

1. Making care safer by reducing harm caused in the delivery of care.
2. Ensuring that each person and family are engaged as partners in their care.
3. Promoting effective communication and coordination of care.
4. Promoting the most effective prevention and treatment practices for the leading causes of mortality, starting with cardiovascular disease.
5. Working with communities to promote wide use of best practices to enable healthy living.
6. Making quality care more affordable for individuals, families, employers, and governments by developing and spreading new health care delivery models.

Access the National Quality Strategy using this link: http://www.ahrq.gov/workingforquality/.

[1] http://www.ahrq.gov/workingforquality/nqs/nqs2012annlrpt.pdf

Measure contractors

Measure contractors should be knowledgeable of the National Quality Strategy aims and priorities and understand how the measures they are developing and maintaining align with these strategies.

Public and stakeholder input

CMS conducts its measurement activities in a transparent manner; this includes posting the Blueprint to inform the public and stakeholders about the processes that CMS contractors use to develop and maintain quality measures. Access the Blueprint using this link: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/index.html.

The opportunities for public and stakeholder input include:

1. Posting calls on the CMS Web site for nominations for technical expert panels (Refer to the Technical Expert Panel section).
2. Posting the proposed or candidate measures on the CMS Web site for public comment and posting the comments that were received (Refer to the Public Comment section).
3. Publishing the proposed and final rules in the Federal Register.
4. Holding CMS Open Door Forums.
5. The National Quality Forum (NQF)-convened MAP (described in greater detail below).

Measure contractors

Measure contractors are expected to be knowledgeable of the inputs into the measurement activities and at a minimum follow the Blueprint processes for posting calls for technical expert panels, and posting candidate measures for comment. (Refer to the Technical Expert Panel and Public Comment sections for further details.) Contractors are responsible for monitoring all feedback and input provided on their measure. It is their responsibility to report this information to their Contracting Officer Representative/Government Task Leader (COR/GTL), who will in turn ensure that CMS staff members working on measures priorities planning receive this information.

Measure maintenance

Once a measure is in use, it requires periodic reevaluation to determine whether its strengths and limitations related to the measure evaluation criteria have changed since the last formal evaluation (Refer to Volume 2 of the Blueprint for measure maintenance information).

Measure contractors

Measure contractors must convey to their COR/GTL—the lessons learned from the measure rollout, the environmental scan, and ongoing monitoring of the measures. The COR/GTL will then share this information with CMS staff leading the measures priorities planning tasks. Measure monitoring may result in information that CMS leadership may find valuable for setting priorities and planning future measurement projects. CMS may request an evaluation of current measures and sets used in the program/initiative and recommendations for ways to accommodate cross-setting use of the measures. The evaluation may also include options for alternative ways to interpret the measures and measure

sets through the continuum of care. Providing this type of feedback closes the loop of the CMS quality development and selection activities.

Reports and analyses

The quality performance of the measures is analyzed by a variety of organizations and entities and these results can provide input into CMS measure priority planning.

Measure contractors

Measure contractors must maintain their measure by reevaluating them every three years and conducting ad hoc reviews as necessitated. This is done to determine whether its strengths and limitations related to the measure evaluation criteria have changed since the last formal evaluation. Volume 2 of the Blueprint provides measure maintenance information to assist in these processes.

Measure contractors should monitor and analyze the following related to their measures:

- Overall performance trends
- Variations in performance
- Gaps in care
- Extent of improvement
- Disparities in the resulting rates by race, ethnicity, age, socioeconomic status, income, region, gender, primary language, disability, or other classifications
- Frequency of use of exclusions or exceptions and the impact on the rates
- Patterns of errors in data collection or rate calculation
- Gaming or other unintended consequences
- Changes in practice that may adversely affect the rates
- Impact of the measurement activities on providers

In addition to conducting their own analyses, measure contractors should also be aware of and monitor other entities that analyze CMS measure performance. Some of these entities and their associated reports are identified below.

- MedPAC—quality reports
- AHRQ—*National Healthcare Quality and Disparities Reports*
- CMS Center for Strategic Planning—*Chronic Conditions among Medicare Beneficiaries*
- Universities and health care facilities—journal articles, conference presentations

2.3 Quality Measure Development and Endorsement Projects

Volume 1 of the Blueprint describes the standardized approach for developing the quality measures CMS uses in its quality initiatives and programs. The Blueprint is composed of a set of business processes and decision criteria that CMS-funded measure contractors are expected follow when developing, implementing, and maintaining quality measures.

Measure contractors

Measure contractors should consider HHS and CMS goals, legislative mandates, public and stakeholder input, NQS priorities, and measure maintenance reports and analyses when identifying a list of potential measures for pre-rulemaking, rulemaking, and eventual program adoption.

To the extent possible, CMS uses NQF-endorsed measures in its programs. The NQF's measure endorsement process is standardized in a regular cycle of topic-based measure evaluation. The measure contractors are responsible for completing the NQF Measure Submission Form and ensuring that the information is sufficient to meet NQF's requirements. The Measure Submission Form is the developer's presentation of the measure to the Steering Committee and others to demonstrate that the measure meets the criteria for endorsement. The measure contractors are also responsible for supporting CMS in answering questions about the measures to the NQF Steering Committees. Access the NQF Web site using this link: http://www.qualityforum.org.

2.4 Pre-Rulemaking Process

Section 3014 of the ACA requires the establishment of a federal pre-rulemaking process for the selection of quality and efficiency measures for specific qualifying programs within HHS, and consideration of multi-stakeholder input on the selection of quality and efficiency measures prior to rulemaking. To meet these requirements, HHS develops a Measures Under Consideration (MUC) list in which the NQF Measures Application Partnership provides input as to the best measures for use in a given program. HHS must consider MAP input and publish the rationale for selecting any performance measures—in proposed or final rules—not endorsed by NQF.

Measures Under Consideration

Over the past few years, CMS has articulated a number of measure selection criteria in its Federal Rules for various programs. The term measure selection typically applies to selection of a set of measures while measure evaluation applies to evaluating an individual measure. In order for HHS to develop the MUC list for qualifying programs publicly available by December 1 of each year, a set of measure selection criteria has been established by CMS. These selection criteria are operationalized by program staff through the CMS Quality Measures Task Force that reviews each measure for program adoption.

Core criteria

- Measure addresses an important condition/topic with a performance gap and has a strong scientific evidence base to demonstrate that the measure when implemented can lead to the desired outcomes and/or more appropriate costs (i.e., NQF's importance criteria).
- Measure addresses one or more of the six National Quality Strategy Priorities (safer care, effective care coordination, preventing and treating leading causes of mortality and morbidity, person- and family-centered care, supporting better health in communities, making care more affordable).
- Promotes alignment with specific program attributes and across CMS and HHS programs.

- Program measure set includes consideration for health care disparities.
- Measure reporting is feasible.

Optional criteria

- The measure is responsive to specific program goals or requirements.
- Enables measurement using a measure type not already measured well (for example structure, process, outcome, etc.).
- Enables measurement across the person-centered episode of care, demonstrated by the assessment of the person's trajectory across providers and settings.
- Measure set promotes parsimony.

CMS develops the MUC list after receiving the measure information from their contractors. CMS then provides this list is to the MAP.

Measure contractors

The measure contractors will assist CMS by providing information about their measures for presentation to the MAP. Contractors may also be required to help the COR/GTL develop the Measures Under Consideration list, which could include helping CMS to evaluate their measures by applying the selection criteria.

MAP recommendations

HHS contracted with NQF, a consensus-based entity, for the specific purpose of convening multi-stakeholder groups—the Measure Applications Partnership. The MAP provides input to HHS, by February 1 of each year, on the identification of the best available performance measures for use in specific programs. For the 2011-2012 cycle, the MAP provided input on over 350 measures under consideration. The MAP provided HHS with three general categories of feedback for each measures reviewed, which include:

- Support the measure—MAP supported the measure for inclusion in the associated federal program during the most current rulemaking cycle for that program. (Approximately 40 percent of the measures under consideration).
- Support the direction of the measure—MAP supported the measure concept, however, further development, testing, or implementation feasibility must be addressed before inclusion in the associated federal program. (Approximately 15 percent of the measures under consideration).
- Do not support the measure—Measure was not recommended for inclusion in the associated federal program. (Approximately 45 percent of the measures under consideration. For nearly 70 percent of the measures within the do not support category, MAP did not have enough information to complete its evaluation, so could not support those measures at the time).[2]

[2] National Quality Forum. Measure Applications Partnership. Pre-Rulemaking Report: Input on Measures Under Consideration by HHS for 2012 Rulemaking. Final Report. February 2012.

The feedback categories are subject to change as the MAP continues to refine how they evaluate measures/sets of measures and how that input is submitted.

In addition to providing input to HHS on measures under consideration for nearly twenty federal programs, the MAP developed a framework for aligning performance measurement that is used to support its decision making and explains why alignment is important. The MAP work represents the first time a public-private partnership has worked together in advance of federal health care rulemaking to provide upstream input on the optimal measures for use in particular programs.

Access the MAP Web pages on the NQF Web site using this link: http://www.qualityforum.org/Setting_Priorities/Partnership/Measure_Applications_Partnership.aspx.

2.5 CMS Considers MAP Input for Final Selection

After CMS receives the MAP input, a deliberation process begins to determine which measures will be included in the federal rulemaking processes. The measure selection criteria used during the development of the MUC list, and identified above, are the same criteria used for federal rulemaking. HHS divisions must consider MAP input and publish the rationale for selecting any performance measures—in proposed or final rules—not endorsed by NQF.

2.6 CMS Rulemaking Processes

After the pre-rulemaking process and CMS' selection of measures for possible inclusion in rulemaking, the next steps in the cycle are:

1. Proposed rule—CMS writes the proposed rule, and publishes it the Federal Register. The proposed rule is generally available for public comment for 60 days.
2. Final rule—CMS considers the comments that were received and publishes the final rule in the Federal Register.

Measure contractors

Measure contractors may be asked to assist CMS with the rulemaking processes. During the proposed phase of rulemaking, the contractors will monitor the comments that are submitted on the measures and begin drafting responses for CMS. For the final rule, contractors may also be asked to provide additional documentation on their measures.

2.7 Rollout, Production, and Monitoring of Measures

Lessons learned from the measure rollout, production, and ongoing monitoring of the measure should be conveyed to the CMS staff leading the measures priorities planning task. CMS may request an evaluation of current measures and sets used in its programs and initiatives and recommendations for ways to accommodate cross-setting use of the measures. Refer to the Measure Rollout and the Measure Production and Monitoring sections for details on these processes.

2.8 Measures Maintenance and Impact Assessment

Measure maintenance

Measure contractors

Measure contractors conduct their measure maintenance activities as directed by the COR/GTL. Volume 2 describes the ad hoc, annual, and comprehensive maintenance review processes. As a result of the reviews there are five different outcomes for CMS measures when undergoing maintenance review:

- Retire—Cease to collect or report the measure indefinitely. This applies only to measures owned by CMS. CMS will not continue to maintain these measures. (When retiring a measure from a set, consider other measures that may complement the remaining set as a replacement.)
- Retain—Keep the measure active with its current specifications and minor changes.
- Revise—Update the measure's current specifications to reflect new information.
- Suspend—Cease to report a measure. Data collection and submission may continue, as directed by CMS. (This option may be used by CMS for "topped-off" measures where there is concern that rates may decline after data collection or reporting ceases.)
- Remove—A measure is no longer included in a particular CMS program set for one or more reasons. This does not imply that other payers/purchasers/programs should cease using the measure. If CMS is the measure steward and another CMS program continues to use the measure, CMS will continue maintaining the particular measure. If another entity is the steward, the other payers/purchasers/programs that may be using the measure are responsible for determining if the steward is continuing to maintain the measure.

Impact assessment

Also mandated by Section 3014 of the ACA, once every three years, the Secretary must provide a publicly-available assessment of the impact of quality measures. The first CMS Measure Impact Assessment report was published in March 2012. Detailed findings for the implemented measures in eight CMS programs are included. For each program, data which illustrates the trends over time are reported and examines the extent to which many of the measure trends have declined, remained unchanged, or increased. This report can be accessed by using this link: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/QualityMeasurementImpactReports.html.

2.9 How the Measure Manager can assist CMS

The Measures Manager role is to research and consider a wide variety of measure-related information and materials to assist CMS in prioritizing measure development activities. This may include

- Reviewing HHS and CMS strategic plans, goals, and initiatives.

- Monitoring the progress of CMS measure development and maintenance projects against the National Quality Strategy.
- Producing harmonization reports.
- Researching the priorities of key external stakeholders to determine which measure development and maintenance activities are occurring; reviewing legislative mandates; and proposed and final federal rules.
- Supporting various internal CMS and interagency workgroups, and assessing the impact of CMS quality measures.
- Conducting informational seminars on the Blueprint. Each month, the Measures Manager selects a topic or two from the Blueprint to present to CMS and the measure contractors. These sessions are recorded via WebEx and are available for review.
- Providing a handbook for the COR/GTLs to use as a guide that describes roles and responsibilities in managing the tasks of the MMS Blueprint. The handbook is an aid used to coordinate measurement contract task progress and risks.

2.10 How the Measures Manager can assist the Measure Contractors

Though the CMS measure contractors work within the parameters set by their COR/GTLs, their work will likely also be framed by the National Quality Strategy three aims and six priorities. In working with the measure contractors, the Measures Manager is prepared to:

- Provide and maintain the CMS measures inventory to help the contractors in their search for similar measures and potentially identifying opportunities for measure harmonization and alignment across settings. The CMS Measures Inventory is updated quarterly and includes a wide array of measures. Based on status or year of anticipated use, the measures are separated into four categories:
  - Current measures are those that are in use by CMS in various programs and settings.
  - Future measures are those that CMS is considering using within the next few years.
  - Retired measures are fully developed measures that CMS no longer uses for a variety of reasons.
  - Archived measures are those measures, measure topics, or measure concepts that were considered for use at one time by CMS but not actually used.
- Provide results of monthly journal scans that include measure related topics. The purpose of the journal scan is to provide CMS and measure contractors with up-to-date information on key journal articles that may impact CMS quality reporting programs and measures. The Measures Manager surveys eleven major medical journals each month, summarizes the articles, and comments on their relevance to CMS measures and programs across settings.
- Provide pre-rulemaking process support for CMS in the development of the MUC list.
- Provide harmonization and alignment reports.
- Review draft documents such as deliverables described in the Blueprint and draft NQF submission forms. The Measure Manager's role is to provide technical assistance and to give feedback on the documents' completeness or clarity. In addition, if approved by the COR/GTL, the Measures Manager may be able to provide examples of other measure contractors' work.

  - The COR/GTL is responsible to approving the measure contractor's final products; the Measures Manager role is to provide review and feedback as requested.
- Conduct informational seminars on the Blueprint. Each month, the Measures Manager selects a topic or two from the Blueprint to present to CMS and the measure contractors. These sessions are recorded via WebEx and are available for review.
- Provide technical assistance regarding Blueprint processes on an as needed basis.
- Provide guidance on the NQF endorsement and maintenance processes.

3. Measure Development

3.1 Introduction

The Measure Development section is intended to provide guidance to a set of standardized processes for the development of high caliber measures suitable for submission to the National Quality Forum (NQF) for endorsement. In some cases, the reader is directed to the appropriate sections of this Blueprint for more detailed instructions. The Blueprint sections necessary for full measure development are all found in Volume 1, and outlined below:

- Section 1: Introduction
- Section 3: Measure Development
- Section 4: Measure Evaluation
- Section 5: Harmonization
- Section 6: Information Gathering
- Section 7: Technical Expert Panel
- Section 8: Technical Specifications
- Section 9: eMeasure Specifications
- Section 10: Special Topics
- Section 11: Risk Adjustment
- Section 12: Public Comment
- Section 13: Measure Testing
- Section 14: National Quality Forum Endorsement
- Section 15: Measure Rollout
- Section 16: Glossary

Figure 3-1 shows an overall flow of the processes involved in developing measures from the time the initial contract award until the measure is ready for implementation—the Centers for Medicare & Medicaid Services (CMS) may accomplish this through one or more contracts.

Figure 3-1 Flow of Measure Development Processes

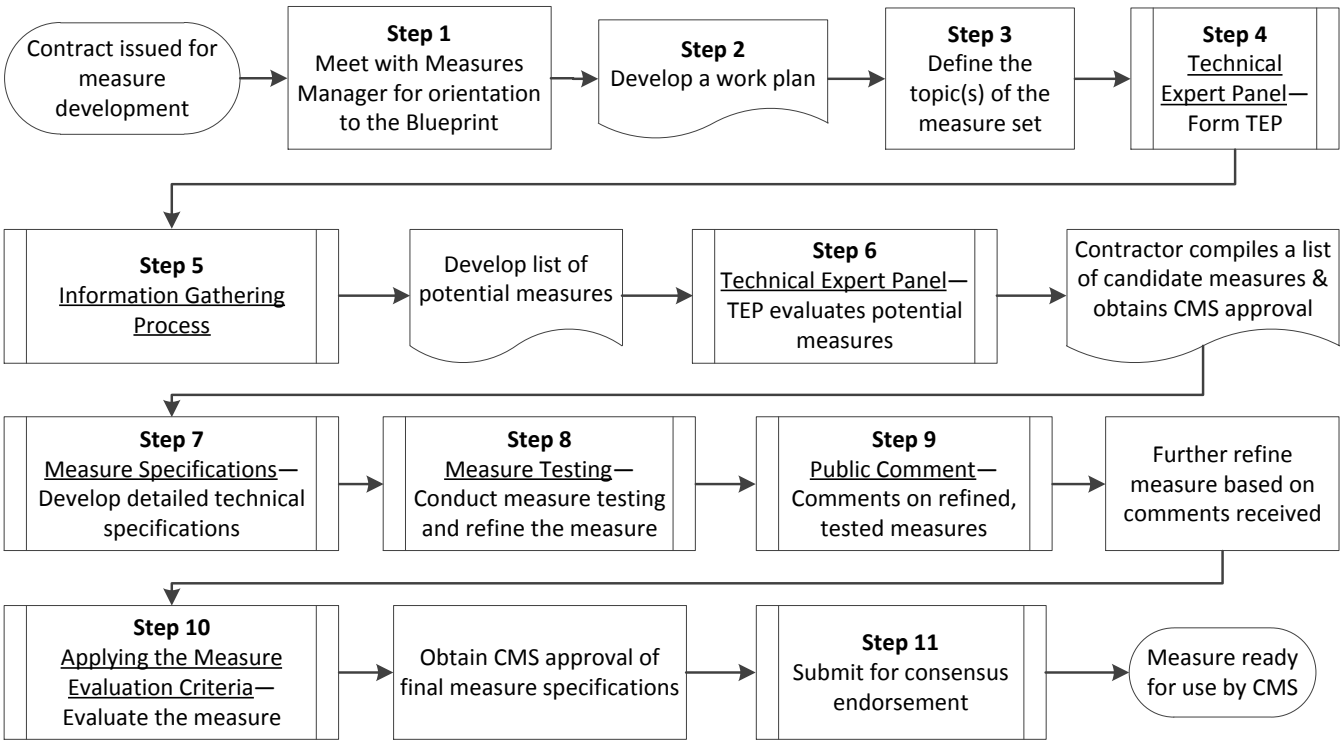
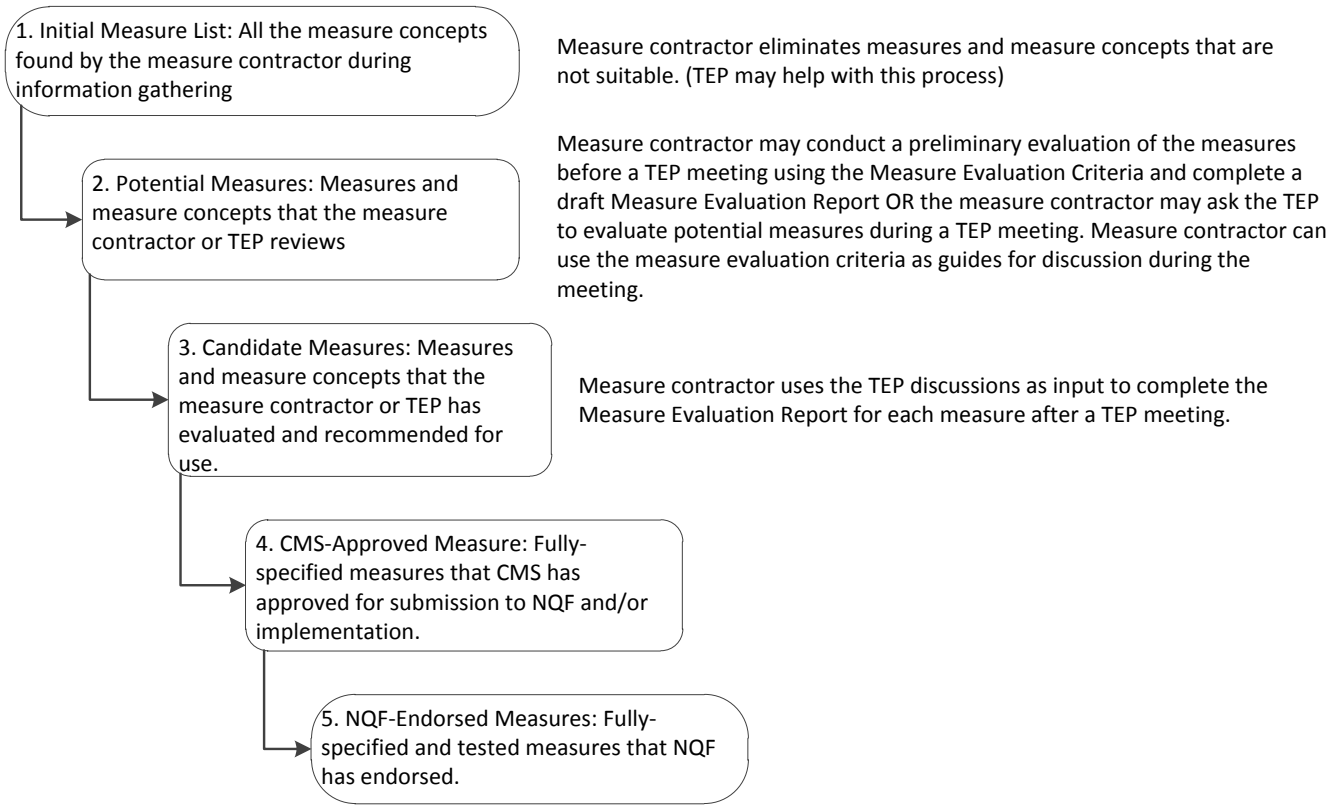


Figure 3-2 depicts how the initial list of measures is refined throughout the measure development process.

- The initial measures list created during information gathering is reviewed and narrowed down to become the list of potential measures evaluated by the technical expert panel (TEP) using the measure evaluation criteria.
- The TEP then forwards the measures it recommends as candidate measures to the measure contractor, who then proposes the candidate measures to CMS.
- Once CMS approves the measures list, detailed measure specifications are documented, and subsequently tested. The measures are then further refined and approved by CMS.
- This smaller list containing newly developed or adapted measures may be submitted to NQF for endorsement.
- NQF may endorse all, some, or none of the measures, based on the measure evaluation criteria.

Figure 3-2 Narrowing Down the Measure List



The end product of measure development is a precisely specified, high-caliber measure to aid CMS in achieving its quality goals. The precisely specified measure is documented in a Measure Information Form (MIF) and Measure Justification form, to allow others to understand the details and rationale of the measure, and allow for consistent interpretation and implementation. These forms are updated incrementally as new information is found and as new and better ways of constructing the measure are identified.

The purpose of measure justification is to provide background material explaining why the measure is important. Measure justification can be thought of in terms of a legal argument: why is this particular measure good enough to expend the resources necessary for data collection, calculation, and reporting? As the measure's specifications are refined, the contractor documents the rationale for any changes on the Measure Justification form. This form is aligned with the NQF Measure Submission Form and can be used to populate the sections relating to the measure evaluation criteria in the NQF Measure Submission Form when the measure is ready to be submitted for endorsement. Consider the following when finalizing measure justification:

- Provide any pilot test data available.
- Identify all possible endorsement roadblocks in advance and address them.
- Document the rationale for all decisions made in the specifications.

- Identify all related and competing measures and describe harmonization achieved or rationale why it was not done.
- Discuss any controversies about the science behind the measure and why the measure was built as it was.
- Document the quality, quantity, and consistency of the evidence supporting the measure.
- Document the case for the measure's reliability and validity.
- Document the rationale for all the measure exclusions or exceptions.
- Document the case for the measure's usability and the costs or burdens of implementation.

Measure justification may not be necessary if the measure being adapted or adopted has already been justified for use in a similar CMS program or has previously been endorsed by the NQF. However, additional justification may be needed to explain why the measure is being modified (adapted) for a new program or setting. The Contracting Officer Representative/Government Task Leader (COR/GTL) will decide if measure justification can be waived for a particular measure.

3.2 Deliverables

Most of the stages of measure development have specific deliverables required by the Blueprint. While these deliverables are important, the end products of measure development are the following documents:

- Completed Measure Information Form (MIF)
  - For contractors developing eMeasures, the Health Quality Measures Format (HQMF) document—which includes a header and a body—should be used in lieu of the MIF. Refer to the eMeasure Specifications section for further details on this format.
- Completed Measure Justification form (if applicable)
- NQF Measure Submission Form (if directed by the COR/GTL)

The MIF and Measure Justification form are included at the end of this section. The electronic NQF Measure Submission Form is available at http://www.qualityforum.org.

3.3 Types of Measures

There are many types of measures. The program measure set should be evaluated for an appropriate mix of measure types, and more and more, CMS is working to ensure that more outcome measures are available across its programs. For the Blueprint, "type of measure" or "measure type" will refer to the following typology.

- Access measure—A measure that focuses on a patient or enrollee's attainment of timely and appropriate health care.
- Composite measure—A combination of two or more individual measures in a single measure resulting in a single score.
- Efficiency measure—A measure of cost of care associated with a specified level of quality of care.

- Outcome measure—A measure that assess the results of health care that are experienced by patients—patients' clinical events, patients' recovery and health status, patients' experiences in the health system, and efficiency/cost. CMS is interested in moving towards greater number of outcome measures rather than process measures.
- Patient experience measure—A measure that focuses on a patient's or enrollee's report concerning observations of and participation in health care.
- Process measure—A measure focusing on a clinical process which leads to a certain outcome, meaning that a scientific basis exists for believing that the process, when executed well, will increase the probability of achieving a desired outcome.
- Cost and resource use measure—Refers to broadly applicable and comparable measures of health services counts (in terms of units or dollars) applied to a population or event (broadly defined to include diagnoses, procedures, or encounters).
- Structural measure—A measure that focuses on a feature of a health care organization or clinician relevant to its capacity to provide health care.

All of these types of measures are developed using the same basic work flow. The procedure included below is intended to guide that work flow. Some types of measures require a modified work flow, and the Special Topics section has been developed to aid in those modifications.

A health quality measure (or clinical quality measure) encoded in the Health Quality Measures Format (HQMF) is referred to as an "eMeasure." eMeasure is an HL7 standard for representing a health quality measure as an electronic XML document. eMeasures are specified in such a way that they use patient-level coded information coded that can be extracted in a format that can be used in the measure. An eMeasure may be a process, outcome, or other type of measure as listed above. eMeasures have special considerations for measure development, as documented in the eMeasure Specifications section.

3.4. Procedure

Step 1: Meet with the Measures Manager for orientation to the Blueprint

A kick off meeting between the measure contractor, the Measures Manager, and the COR/GTL will help ensure that the measure contractor understands the Measures Management System Blueprint prior to measure development. In addition, this is the Measures Manager's opportunity to explain the types of technical assistance it can provide (for example, review documents for completeness, especially the Measure Information Form and the Measure Justification form, the NQF submissions, environmental scan report, assist in identifying related or competing measures, etc.). The meeting may be conducted by conference call, and should be completed within the first two weeks after the contract award.

Step 2: Develop a work plan

As directed by the measure contractor's scope of work, a work plan for the measures to be developed, maintained, or updated will reflect the Measures Management System processes and will provide the

COR/GTL with evidence that the measure contractor understands the processes and has a strategy for executing them. The measure contractor will refer to their contract scope of work for the date when the work plan is due.

When developing the work plan, the measure contractor will consider the schedule for any applicable rulemaking procedures, such as the Inpatient Prospective Payment System (IPPS) or Outpatient Prospective Payment System (OPPS), the processes outlined in ACA Section 3014 for obtaining stakeholder input for the Measures Under Consideration list, and the NQF review cycle.

Step 3: Define the topic(s) of the measure set

The specific measurement topics may be defined by the contract or by the measure contractor as he or she identifies priority topics within the measurement area of interest. The Information Gathering, Technical Specifications, and Measure Evaluation Criteria sections provide guidance on how to proceed with this task. This work should be conducted in close consultation with the COR/GTL.

Step 4: Recruit the Technical Expert Panel

A TEP is a group of stakeholders and experts who provide direction and thoughtful input to the measure contractor on the development and selection of measures for which the contractor is responsible. Convening the TEP is one of the important steps in the measure development process. If the TEP is to be held early during the contract period, then the process of posting the call for nominations should begin as soon as the contract is awarded. This can be done simultaneously with the information gathering and other tasks that are being undertaken at the beginning of the contract. Refer to the Technical Expert Panel section for the standardized selection process.

Step 5: Information gathering process

Refer to the Information Gathering section for further details on the information gathering process outlined below.

Determine the appropriate basis for measures

Based on the material gathered in the Information Gathering process—including clinical guidelines—and in consultation with the TEP, determine the appropriate basis for measures. The appropriate basis will vary by type of measure.

Develop a framework for measures

Based on the material gathered during the Information Gathering process—including clinical guidelines—and in consultation with the TEP, develop a framework for measures. The framework may be based on a typology of measures, with an indication of the types of measures already developed and used extensively (both by CMS and by others). The framework may also organize the existing measures by settings, the National Quality Strategy aims and priorities, by the Institute of Medicine (IOM) aims, by the goals set by CMS, National Priorities Partners (NPP), the Measure Applications Partnership (MAP), or others. The goal of the framework is to identify gaps in existing measures that may require development of new measures.

Develop a business case

Based on the material gathered during the information gathering process, develop a business case for each of the candidate measures.

Search for initial measures

Initial measures should correspond to the draft framework. The search for initial measures includes a search for:

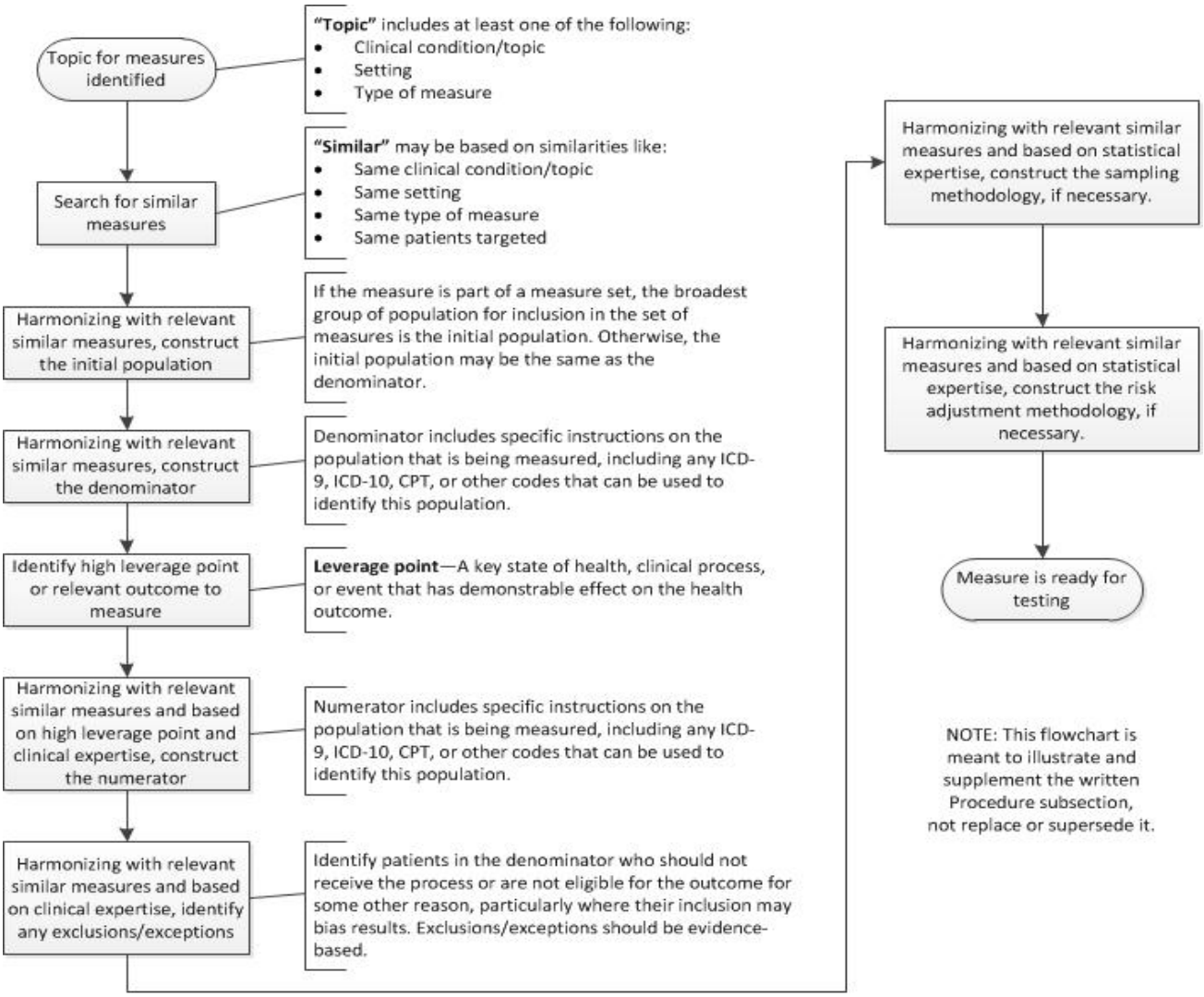
- Relevant clinical guidelines.
- Existing measures that can be adopted (used without change) or adapted (used after some changes have been made).
- Related measures that can serve as models for new measures.
- Studies that can be used as the evidence for new measures.
- Results from a "Call for Candidate Measures," if one has been directed by the COR/GTL.

The initial measures are assessed informally against the measure evaluation criteria, focusing primarily on the importance criterion. (Refer to the Measure Evaluation section). Existing "similar" measures are assessed for possible adoption or adaptation. The TEP may be asked to help with this informal process. Usually the assessment process results in the elimination of some of the initial measures leaving the potential measures (Figure 3-2) to be evaluated more formally by the TEP.

Compile a list of potential measures

Document any candidate measures found using high-level statements in the Measure Information Form (MIF) and the Measure Justification form (if needed). The high-level statements should include tentative descriptions of the proposed denominator and numerator, as well as material justifying the selection of the measure (if needed).

Figure 3-3 Development of a New Measure



If the information gathering process produces similar measures that can be adopted or adapted, please refer to the Information Gathering section. If the information gathering process does not produce sufficient measures (in quantity, caliber, or characteristic), construct high-level measure statements for potential measures using the clinical guidelines, appropriate measure bases, and other information that was found. Refer to the Technical Specifications section for the standardized process. Figure 3-3 above illustrates the development process that can be used when developing a de novo measure.

During this step, also identify measures that present opportunities for harmonization. Harmonization may relate to numerator, denominator, exceptions/exclusions, definitions, and methodology.

Step 6: The TEP evaluates the potential measures

Before the TEP meeting, measure contractor may send a list of potential measures to the TEP and provide supporting rationale as well as any outstanding controversies about the measures. Depending on the specifics of the measure contract, measure contractor may seek TEP guidance on one or more measure evaluation criteria based on TEP expertise and as deemed appropriate by the measure contractor. Please refer to the Measure Evaluation section for detailed instructions. The measure contractor then uses the TEP discussions as input to complete the Measure Evaluation Report for each measure after the meeting. Alternatively, the measure contractor may conduct a preliminary evaluation of the measures and complete a draft Measure Evaluation Report before the TEP meeting. These drafts can be presented to the TEP for discussion.

Once the TEP has made its recommendations to the measure contractor, the contractor will develop a list of candidate measures for submission to the COR/GTL, accompanied by a Measure Evaluation Report for each measure. Refer to the Measure Evaluation section for the standardized reporting process. Though the contractor's list will usually reflect the TEP's recommendations, it is important to note that the recommendations made to CMS come from the contractor, not the TEP.

Obtain the COR/GTL's approval of the recommended measure list

The list of recommended measures will be submitted to the COR/GTL for review. If the measure will be submitted to NQF for endorsement and NQF is using the 2-Stage Consensus Development Process (CDP) for an appropriate topic, the measure contractor will complete the measure concept submission form and obtain COR/GTL approval prior to submission. During Stage 1 of this process, NQF will evaluate the measure concept, focusing on Importance. CMS may instruct the measure contractor to await the NQF approval of the concept before proceeding with the measure's further development and testing.

Step 7: Develop detailed technical specifications using the Measure Information Form and the Measure Justification form

Refer to the Technical Specifications and Measure Testing sections for the standardized specification development processes. Note that measure specifications are developed in an iterative process with alpha (formative) testing and that recruiting facilities or providers to serve as test sites early in the process can help avoid delays in measure development.

For contractors developing eMeasures, the Health Quality Measures Format (HQMF) document—which includes a header and a body—should be used in lieu of the MIF. Refer to the eMeasure Specifications section for further details on this format.

Obtain the COR/GTL's approval of the detailed technical specifications

The MIF will be submitted to the COR/GTL for review. For eMeasures, the measure developer will use the Health Quality Measures Format document.

Step 8: Conduct beta (or field) testing

In order to receive endorsement from the NQF, the measures must be tested. Recruiting facilities or providers to serve as test sites early in the process can help avoid delays in measure development. Refer to the Measure Testing section for the standardized testing process.

Step 9: Solicit public comment

The public comment periods for proposed rules may satisfy this requirement, if the measures will be proposed in any federal rulemaking procedures. The COR/GTL will make this decision.

Once the measures have been beta (or field) tested, conduct a public comment period as directed by the COR/GTL. Refer to the Public Comment section for the standardized comment solicitation process.

Once the comment solicitation period has ended, the comments will be analyzed and reviewed by the measure contractor. If the comments indicate a need to refine the measure, the measure contractor will consult with the TEP and revise the measure as needed. Review the Measure Justification form and update it as necessary.

Step 10: Apply the measure evaluation criteria and evaluate the measures

Once the measures have been fully tested and refined based on both testing and public comment, the contractor updates the Measure Evaluation Report and Measure Justification form. The TEP may be helpful in this process. Refer to the Measure Evaluation and the Technical Expert Panel sections for the standardized processes.

Obtain the COR/GTL's final approval of the measure

The final measure specifications, including testing results and public comments which have been addressed, are submitted to the COR/GTL for review and approval.

Step 11: Submit the measure for NQF consensus endorsement

The COR/GTL will determine readiness for submission to NQF for endorsement. Review the NQF submission requirements on the NQF Web site (http://www.qualityforum.org), including directions on completing the online submission form. Every effort has been made to ensure that the MIF and Measure Justification are aligned with the NQF Measure Submission Form to simplify the process of populating the NQF data fields. Refer to the National Quality Forum section for further information.

NQF will accept notification from measure developers that they have a measure that is ready, or almost ready, to submit and will allow measure developers to begin the online Measure Submission Form at any time, unrelated to a particular NQF project. The purpose of providing the online submission form unrelated to a particular NQF project is to give measure developers another venue to indicate readiness to NQF, regardless of topic. Completing the Measure Submission Form prior to an NQF project allows CMS and the measure contractor to have time to thoroughly review the submission to ensure the information is presented in a complete and thoughtful manner. The Measures Manager can assist in this review.

Prior to submitting the measure to NQF, conduct another search for potential competing measures and measures that have opportunities for harmonization. Although this should have been done early during the information gathering stage of measure development, new information may be available, and should be identified prior to NQF submission.

NQF will convene a steering committee to evaluate measures submitted. The outcomes of this evaluation include:

- Endorsement
- Time-limited endorsement—measures meeting NQF requirements except field testing
- Deferred endorsement—pending further information from the measure contractor
- Declined endorsement

Meet to review the NQF endorsement results

After the NQF has completed its CDP, the measure contractor will meet with the contractor's COR/GTL and the Measures Management Team to discuss the results of the CDP, why CMS measures were not endorsed (if any were not), identify lessons learned regarding both the NQF process and CMS Measures Management System processes.

Measure Name De.1. *Briefly convey as much information as possible about the measure focus and target population—abbreviated description*

3a Measure Information Form (MIF)

Data Source

- 2a1.25-Data Source
- 2a1.26-Data Source or Collection Instrument
  Identify the specific data source or data collection instrument (for example, name of database, clinical registry, collection instrument, etc.)

Measure Set ID

- This field can be used to describe any ID or numbering used for the set of measures. This may be left blank if it is not applicable

Version Number and effective date

- Similar to Ad.4

CMS approval date

- The date when CMS approved this version of the measure

NQF ID

- The number assigned by NQF after endorsement. If NQF assigns and interim ID, this ID can be used, but should be updated when permanent number is assigned.

Date Endorsed

- This date should be most recent endorsement date and match the date on the NQF site

Care Setting

- 2a1.34

Unit of Measurement

- Level of Analysis: 2a1.33

Measurement Duration

- NQF uses Numerator(2a1.2) and Denominator(2a1.6) **Time Windows**

Measurement Period

- Year, month, quarter, etc.

Measure Type

- (included in “Importance” 1c.1)

Measure Scoring

- Type of Score: 2a1.17

Payer source

- Describe the payer for cases included in the measure. (for example Medicare only, all payers, Medicare and Medicaid)

Improvement notation

- Interpretation of Score: 2a1.17

Measure steward

- CO1.1

Copyright / Disclaimer

- Please state any copyright or disclaimer that the owner may require. This does not apply to measures developed and owned by CMS.

Measure description

- De.2. Briefly describe the type of score (for example, percentage, proportion, number) and
- the target population and focus of measurement:

[type of score] of [target population] who received/had [measure focus]

Rationale

- Succinct statement of the need for the measure. Usually includes statements pertaining to Importance criterion: impact, gap in care and evidence.

Clinical Recommendation Statement

- 1c16 (This is also included in the complete Measure Justification section) This information is repeated here so that the MIF can be used as a stand-alone document. This field is used to state the wording from the guideline.

References

- 1c17 Provide references for the guideline used or studies if there is not a guideline.

Release Notes / Summary of Changes

- Describe changes to measure if this is not the original version

Technical Specifications

- Target Population

  NQF form 2a1.5, "Target Population Category" does not match the intent of this field. The intent of this field was to approximate the e-measure filed "Initial Population." NQF uses the term "Target Population" to mean "denominator."

Denominator

- Denominator Statement

  2a1.4. Designate the broadest population based on the evidence for which the target process, condition, event, and outcome is applicable. The target population should indicate age, setting, and time frame for identifying the target population.

Patient's [age] with [condition] in [setting] during [time frame]

- Denominator Details
- *2a1.7.*

  Codes: For measures based on a coded data set, identify the code set, the specific codes, and descriptors for the codes.
  Details: Definitions and instructions as needed.

- Denominator Exceptions and Exclusions

  *2a1.8. (**NQF includes "exceptions"** in the "exclusion "field)*
  Identify patients who are in the target population, but who should not receive the process or are not eligible for the outcome for some other reason, particularly if their inclusion may bias results. Exclusions should be evidence-based.

  Patients in the [target population] who [have some additional characteristic, condition, procedure]

- Denominator Exceptions and Exclusions Details

  *2a1.9. (**NQF includes "exceptions"** in the "exclusion "field)*
  Codes: For measures based on a coded data set, identify the code set, the specific codes, and descriptors for the codes.
  Details: Definitions and instructions as needed.

Numerator

- Numerator Statement

  2a1.1. Describe the measure focus—cases from the target population with the target process, condition, event, or outcome based on the evidence.

  Patients in the target population who received/had [measure focus] {during [time frame] if different than for target population}

- Numerator Details

  2a1.3.
  Codes: For measures based on a coded data set, identify the code set, the specific codes, and descriptors for the codes.
  Details: Definitions and instructions as needed.

Stratification or Risk Adjustment

2a1.10. Stratification variables: Provide instructions for calculating the measure by category (for example, age) including the stratification variables, all codes, logic, and definitions.
or
2a1.11, 2a1.13, 2a1.14 Risk Adjustment:

- Identify the method and variables/risk factors (not the details) in this field.
- *Provide risk model coefficients or equation to estimate each patient's probability for the outcome including coefficients for the variables/risk factors. Provide the codes or definitions for each variable/risk factor. Provide programming language (for example, SAS code). It is usually necessary to provide an attachment or URL for these details.*

Sampling

2a1.24. If measure is based on a sample (or survey), provide instructions for obtaining the sample and conducting the survey, and guidance on minimum sample size (response rate).

Calculation Algorithm

Calculation Algorithm/Measure logic: 2a1.20, 2a1.21. Describe the calculation of the measure as a flow chart or series of steps.

Measure Name De.1. *Briefly convey as much information as possible about the measure focus and target population—abbreviated description*

3b Measure Justification

Importance

- ***High Impact Aspect of Health Care***
  - **Demonstrated high impact aspect**

  1a1.1 Select from the following all that apply:
    - *Affects large numbers*
    - *A leading cause of morbidity/mortality*
    - *Frequently performed procedure*
    - *High resource use*

  Patient/societal consequences of poor quality

  - **Summary of evidence of high impact**

  1a3. Provide epidemiological or resource use data

  - **Citations**

  1a.4. Provide citations for the evidence described above

- ***Opportunity for Improvement***
  - **Briefly explain the benefits envisioned by use of this measure**

  1b.1. (Quality improvement anticipated)

  - **Summary of data demonstrating performance gap**

  1b.2. (Variation or overall less than optimal performance across providers)

  - **Citations**

  1b.3. Provide citations for the evidence described above

  - **Summary of data on disparities by population group**

  1b.4.Summarize evidence found that demonstrates any disparities. Describe groups in which disparities exist.

  - **Citations**

  1b.5. Provide citations for the evidence described above

- ***Evidence to Support Measure Focus***
  - **Structure-process-outcome relationship**

  1c.1. Briefly state the measure focus (for example, health outcome, intermediate clinical outcome, process, structure) Then, identify the appropriate links (for example, structure-process-health outcome, process-health outcome, intermediate clinical outcome-health outcome)

  - **Type of evidence**
    - *1c.2. Describe the type of evidence, selecting from the following list all that apply: Clinical practice guideline*
    - *Selected individual studies (rather than entire body of evidence)*
    - *Systematic review of body of evidence (other than within guideline development)*

Other (state type of evidence)

- **Directness of evidence to the specified measure**

1c.4. State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population.

- **Quantity of studies in the body of evidence**

1c.5. Total number of studies, not articles

- **Quality of body of evidence**

1c.6. Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address:

a) Study design/flaws
b) Directness/indirectness of the evidence to this measure (for example, interventions, comparisons, outcomes assessed, population included in the evidence)
Imprecision/wide confidence intervals due to few patients or events)

- **Consistency of results across studies**

1c7. Summarize the consistency of the magnitude and direction of the effect across studies

- **Net benefit**

1c8. Provide estimates of effect for benefit/outcome, identify harms addressed and estimates of effect, and net benefit---benefit over harms across studies. Please include results of business/social/economic case for the measure.

- **Grading of strength/quality of the body of evidence**

1c9, 1c10. 1c11, 1c13, 1c14. Please address:

- *Indicate if the body of evidence has been graded*
- *If the body of evidence was graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias*
- *System used for grading the body of evidence*
- *Grade assigned to the body of evidence*
  Summary of controversy/contradictory evidence

- **Citation**

1c15. Provide citations for the evidence described above

- **Guideline recommendation**

1c16. Quote verbatim, the specific guideline recommendation (Including guideline number and/or page number)

- **Citation**

1c17. Provide citations for the clinical practice guideline quoted above

- **URL**

1c18. National Guideline Clearinghouse or other URL

- **Grading of strength of recommendation**

1c191 1c21, 1c23. Please address:

- *Has the recommendation been graded?*

- *System used for grading the strength of guideline recommendation (USPSTF, GRADE, etc.) Grade assigned to the recommendation*

- **Rationale for using this guideline over others**

1c24. If multiple guidelines exist, describe why the guideline cited was chosen. Factors may include rigor of guideline development, widespread acceptance and use, etc.

- **Overall assessment of the body of evidence**

1c25, 1c26, 1c.27. Based on the NQF descriptions for rating the evidence, what was your assessment of the following attributes of the body of evidence?

- *Quantity*
- *Quality*
- *Consistency*

Reliability and Validity – Scientific Acceptability of Measure Properties

- ***Reliability Testing***
  - **Data sample**

2a2.1.Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included

  - **Analytic methods**

2b2.2 .Describe method of validity testing and rationale; if face validity, describe systematic assessment

  - **Testing results**

2a2.3. Provide reliability statistics and assessment of adequacy in the context of norms for the test conducted

- ***Validity Testing***
  - **Data sample**

2b2.1. Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included

  - **Analytic method**

2b2.2 .Describe method of validity testing and rationale; if face validity, describe systematic assessment

  - **Testing results**

2b2.3. (Provide statistical results and assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)

- ***Exclusions***
  - **Data sample for analysis of exclusions**

2b3.1.Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included

  - **Analytic method**

2b3.2. Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference

  - **Results**

2b3.3. Provide statistical results for analysis of exclusions (for example, frequency, variability, sensitivity analyses)

- ***Risk Adjustment Strategy***
  - **Data/ sample**

    2b4.1. Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Delete row if measure is not risk adjusted.

  - **Analytic method**

    2b4.2. Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables

  - **Testing results**

    2b4.3. Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata. Delete row if measure is not risk adjusted.

  - **Rationale for no adjustment**

    2b4.4. If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment. The three rows above may be deleted if this field is used. Delete row if measure is risk adjusted or if this is a process measure.

- ***Identification of Meaningful Differences in Performance***
  - **Data/ sample**

    2b5.1 Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included

  - **Analytic method**

    2b5.2. Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance

  - **Testing results**

    2b5.3. Results-Provide measure performance results/scores (for example, distribution by quartile, mean, median, SD, etc.); identification of statistically significant and meaningfully differences in performance

- ***Comparability of Multiple Data Sources/Methods***
  - **Data/ sample**

    2b6.1. Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included

  - **Analytic method**

    2b6.2. Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure

  - **Testing results**

    2b6.3. Provide statistical results (for example, correlation statistics, comparison of rankings) and assessment of adequacy in the context of norms for the test conducted

- ***Disparities in Care***
  - **Stratification**

  2c.1. If measure is stratified for disparities, provide stratified results (scores by stratified categories/cohorts)

  - **Rationale for no stratification**

  2c.2. If disparities have been reported/identified, but measure is not specified to detect disparities, please explain.

  - **Supplemental information**

  2.1. Supplemental testing methodology information: If additional information if available, please indicate where this information can be found: appendix, attachment, or URL

Usability

- ***Public Reporting***
  - **Meaningful, understandable and useful**

  3a.1. Use in public reporting---disclosure of performance results to the public at large (If used in a public reporting program, provide name of program(s), locations, Web page URL(s). If not publicly reported in a national or community program, state the reason and plans to achieve public reporting, potential reporting programs or commitments, and timeline, for example, within 3 years of endorsement)
  3a.2. Provide a rationale for why the measure performance results are meaningful, understandable, and useful for public reporting. If usefulness was demonstrated (for example, focus, group, cognitive testing) describe the data, method and results.

- ***Quality Improvement***
  - **Meaningful, understandable and useful**

  3b.1. Use in QI (If used in quality improvement program, provide name of program(s), locations, Web page URL(s))
  3b.2. Provide a rationale for why the measure performance results are meaningful, understandable, and useful for quality improvement. If usefulness was demonstrated (for example, QI, initiative) describe the data, method and results

  - **Other accountability uses**

  3.2. Use for other accountability functions (payment, certification, accreditation) (If used in a public accountability program, provide name of program(s), locations, Web page URL(s)). This row may be deleted if not applicable.

Feasibility

- ***How the data elements needed to compute measure score are generated***

  4a.1. How are the data elements needed to compute measure scores generated? State all that apply. Data used in the measure are:

  - *Generated by and used by health care personnel during the provision of care (for example, blood pressure, lab value, medical condition)*
    Coded by someone other than person obtaining original information (for example, DRG, ICD-9 codes on claims)
  - *Abstracted from a record by someone other than person obtaining original information (for example, chart abstraction for quality measure or registry) Other*

- ***Electronic availability***

  4b.1. Are the data elements needed for the measure as specified available electronically (elements that are needed to compute measure scores are in defined, computer-readable fields)?

- *ALL data elements in electronic health records (EHRs)*
- *ALL data elements in electronic claims*
- *ALL data elements are in a combination of electronic sources (describe)*
- *Some data elements are in electronic sources (describe)*
  - *No data elements are in electronic sources*

- ***Susceptibility to inaccuracies, errors, or unintended consequences***
  4c.1. Identify susceptibility to inaccuracies, errors, or unintended consequences of measurement identified during testing and/or operational use and strategies to prevent, minimize, or detect. If audited, provide results.

- ***Data collection strategy***
  4d.1. Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues (for example fees for use of proprietary measures)

Related Measures

- ***Harmonization***
  5a.1. If this measure has EITHER the same measure focus OR the same target population as NQF-endorsed measure(s): Are the measure specifications completely harmonized? Is so, describe.

- ***Similar measures***
  5b.1. If this measure has both the same measure focus and the same target population as NQF-endorsed measure(s) or other measures in current use, describe why this measure is superior to existing measures (for example, a more valid or efficient way to measure quality); OR, provide a rationale for the additive value of developing and endorsing an additional measure. (Provide analyses when possible.)

4. Measure Evaluation

4.1 Introduction

The Centers for Medicare & Medicaid Services (CMS) aim to develop quality measures of the highest caliber that will drive significant health care quality improvement and inform consumer choices. To gain CMS approval for measure implementation, the measure developer must first provide strong evidence that the measure adds value to existing measurement programs and that it is constructed in a sound manner. CMS gives preference to measures that are already endorsed by the National Quality Forum (NQF) or are likely to become endorsed for implementation in its programs. Therefore, measure contractors should develop measures that meet NQF evaluation criteria and are likely to be endorsed if they are submitted for endorsement.

Each proposed measure undergoes rigorous evaluation during the development process to determine its value and soundness based on a set of standardized criteria including—importance to measure and report, scientific acceptability of the measure properties, usability and use, feasibility, and in certain cases, harmonization. NQF requires measure harmonization as part of their endorsement and endorsement maintenance processes, placing it after initial review of the four measure evaluation criteria. Since harmonization should be considered from the very beginning of measure development, CMS contractors are expected to consider harmonization as one of the core measure evaluation criteria. (Please refer to the Harmonization section for further information). Each criterion is comprised of a set of subcriteria. The measure developer uses an iterative evaluation process to build and strengthen the measure justification and technical specifications to demonstrate the following:

- The aspects of care included in the specifications are highly important to measure and report because the results can supply new, meaningful information to consumers and health care providers to drive significant improvements in health care quality and health outcomes where there is variation in or overall less-than-optimal performance.
- The data elements, codes, and parameters included in the specifications are the optimum choices to quantify the particular measure because they most accurately and clearly target the aspects that are important to collect and report, and they do not place undue burden on resources in order to collect the data.
- The calculations included in the specifications are optimum methodologies because they reflect a clear and accurate representation of the variation in the quality or efficiency of care delivered or the variation in the health outcome of interest.

This section (including its appendices) provides an overview of the measure evaluation criteria and subcriteria, and provides guidance for evaluating measures. A template for the measure evaluation report is included. The Measure Evaluation Report documents for CMS the extent to which the measure meets the criteria. It also documents any plans the measure contractor has to improve the rating, when a measure is rated low or moderate on any sub criterion. The measure contractor will use this report to document the pros and cons, cost benefit, and any risks associated with not further refining the measure. To facilitate efficient and effective development of high caliber measures, the

materials in this section have been revised to reflect changes implemented by the National Quality Forum (NQF) as of January 2012, as well as additional criteria developed by CMS for evaluating a measure set.

4.2 Deliverables

- When recommending approval of *candidate* measures for further development—submit a separate measure evaluation report for each measure.
- When recommending approval of *fully tested and refined* measures for implementation—submit a separate measure evaluation report for each measure.

4.3 Discussion of Measure Evaluation Criteria and Subcriteria

Importance to Measure and Report

High impact area

This criterion emphasizes that the specific measure focus (i.e., what is measured) should be considered important enough to expend resources for measurement and reporting according to CMS goals and priorities as well as recognized national health goals and priorities.

In addition to addressing national priorities, measures may be warranted for some other high impact aspects of health care. The subcriteria in the importance category address many issues related to these high impact aspects of care, such as: leading cause of morbidity/mortality, high resource use, severity of consequences of poor quality, number of people at risk, effectiveness of care, and opportunity for improvement. These aspects should be systematically evaluated by developing a business case for the measure. Not all topics are associated with financial savings to Medicare; however, a topic may be beneficial for society in general or may be of ethical value. The benefits derived from the interventions promoted by the quality measures should be quantified whether in terms of a business, economic or social case. (Refer to the Information Gathering section appendices for more information and examples.)

Opportunity for improvement/gap in care

It is not enough that the measure is merely related to an important broad topic area. Evaluate whether the measure focus is a quality problem (i.e., there must be an opportunity or gap between actual and potential performance). Examples of opportunity for improvement data include, but are not limited to, prior studies, epidemiologic data, and measure data from pilot testing or implementation. If data are not available for review, the measure focus should be systematically assessed (e.g., expert panel rating) and evaluated relative to being a quality problem.

Generally, rare event outcomes do not provide adequate information for improvement or discrimination. However, “never events” that are compared to zero are appropriate outcomes for public reporting and for quality improvement.

Evidence to support measure focus

Health outcomes are often the preferred focus of a measure because they integrate the influence of multiple care processes and disciplines involved in the care. Because multiple processes influence a health outcome, health outcomes generally do not require empirical evidence linking them to a known process or structure of care.[1] For other (non-outcome) types of measures, there must be a high-to-moderate degree of certainty, as demonstrated by the evidence, that the measure focus is linked to positive outcomes. For intermediate outcome measures, process measures and structure measures, evidence should be evaluated on the quantity of studies, the quality of those studies, and the consistency in direction and magnitude of the net benefit of the body of evidence. When clinical practice guidelines are used to support the measure focus, the methodological rigor and review should be understood to ensure that the underlying evidence meets the quality, quantity and consistency requirements.

If the measure focus is a single step in a multi-step process, select the step with the greatest effect on the desired outcome. For example, although assessing immunization status and recommending immunization are necessary steps, they are not sufficient to achieve the desired impact on health status—patients must be vaccinated to achieve immunity. This does not preclude consideration of preventive screening interventions measures where there is a strong link with desired outcomes (e.g., mammography). This also does not preclude selecting a measure focus within a multi-step process that is consistent with the provider's scope of practice.

NQF is currently restructuring its Consensus Development Process (CDP) to include a two-stage process.[2] The proposed two-stage process will incorporate an early review and approval (1st stage) of measure concepts against the importance criterion before it moves to further stages of measure development (i.e. testing and specification). Potential related and competing measures would also be identified during this stage. If a measure is to be submitted for NQF endorsement, the importance and harmonization criteria are considered "must pass" before the measure will be given any further consideration, or allowed to advance to the 2nd stage. The 2nd stage of the CDP allows measure developers 18 months to complete testing and full specification of measures. At this stage, measures will be evaluated for scientific acceptability, feasibility and usability.

Scientific Acceptability of Measure Properties

Scientific acceptability of measure properties addresses the basic measurement principles of reliability and validity. This criterion focuses on the extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. The subcriteria for both reliability and validity reflect this focus.

---

[1]National Quality Forum. *Guidance for Evaluating Evidence Related to the Focus of Quality Measurement and Importance to Measure and Report.* January 2011. Available at: http://www.qualityforum.org/Measuring_Performance/Improving_NQF_Process/Evidence_Task_Force.aspx Accessed: June 2011.

[2]National Quality Forum. *Overview of 2-Stage Consensus Development Process Redesign*. April 2012. Available at: http://www.qualityforum.org/Measuring_Performance/Improving_NQF_Process/Two-Stage_Consensus_Development_Process_Redesign.aspx Accessed: July 31, 2012.

Reliability

Evaluating a measure's reliability requires assessment of the measure's specifications and empirical evidence of the measure's reliability testing. A reliable measure is well-defined and precisely specified; thus, it can be implemented consistently within and across organizations and allow for comparability. Threats to reliability include ambiguous measure specifications (including definitions, codes, data collection, and scoring).

Although precise specifications provide a foundation for consistent implementation and increase the likelihood of reliability, reliability cannot be assumed. Reliability testing demonstrates repeatability of data elements for the same population in the same time period, and precision of performance scores. Therefore, evaluation of a measure's reliability involves an assessment of the empirical evidence of reliability of both data elements used to calculate the measure score and the computed measure score.

Validity

Evaluation of a measure's validity involves an assessment of the consistency between measure specifications and evidence presented to support the measure focus, empirical evidence of a measure's validity testing, and threats to validity.

The evidence for the measure focus as identified in the Importance to Measure and Report criterion provides a foundation for the validity of the measure as an indicator of quality. Therefore, evaluation of a measure's validity entails a review of the measure specifications (numerator, denominator, exclusions, risk factors) and the evidence that supports them. If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure. In such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

In addition, the way a measure is specified can affect the validity of the conclusion about quality. Evaluation of a measure's validity involves an assessment of the results of the empirical evidence of validity of both data elements and measure score.

Evaluation of a measure's validity also involves assessing for the identification and empirical assessment of threats to validity to prevent biased results. Threats to validity include other aspects of the measure specifications such as inappropriate exclusions, lack of appropriate risk adjustment or risk stratification for outcome and resource use measures, use of multiple data sources or methods that result in different scores and conclusions about quality, and systematic missing or "incorrect" data.

With large enough sample sizes, even small differences can be statistically significant. However, they may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025)

is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

Risk adjustment

Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for cardiovascular disease risk factors between men and women). It is preferable to stratify measure results by race and socioeconomic status rather than to adjust out the differences. (Refer to the Risk Adjustment section for discussion of risk adjustment model adequacy testing methods.)

Feasibility

This criterion evaluates the extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement. Feasibility is important to the adoption and ultimate impact of the measure and needs to be assessed through testing or actual operational use of the measures.

Confidentiality is important to feasibility, and is affected primarily by how a measure is implemented rather than the measure specifications. All data collection must conform to laws regarding protected health information. Confidentiality may be a particular issue with measures based on patient surveys or when there are small numbers of patients.

Other feasibility subcriteria address methods to reduce measurement burden such as minimizing exclusions and use of electronic data sources, taking into consideration the status of transition to electronic means for data collection. Ultimately, clinical quality measures should be derived from clinical data as a byproduct of patient care. During the transition period, one approach is to use measures based on clinically enriched electronic administrative data.

As a part of feasibility, NQF requires that susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit for them are assessed. Because NQF's responsibilities do not include implementation of measures, they focus on the purpose of auditing, rather than evaluating the audit strategy, which is considered an implementation issue.

Usability and Use

A measure is evaluated for usability and use after it has met the other three major criteria. When a measure meets the importance, scientific acceptability and feasibility criteria, that particular measure is potentially usable. Evaluation of a measure's usability and use involves an assessment of the extent to which a measure has been used or can be used in accountability applications and in performance improvement. Accountability applications refer to the use of performance results about identifiable, accountable entities to make judgments and decisions as a consequence of performance such as reward, recognition, punishment, payment, or selection (e.g. public reporting, accreditation, licensure, professional certification, health IT incentives, performance-based payment, network

inclusion/exclusion). Selection is the use of performance results to make or affirm choices regarding providers of healthcare or health plans.

Performance results on measures that are in use in an accountability application must demonstrate progress in achieving high quality healthcare. Lack of use or lack of progress in achieving high quality healthcare may be indicative of problems related to the other criteria. A reexamination of the other three criteria may be necessary.

The expectation of being useful for "informing quality improvement" allows consideration of important outcome measure that may not have an identified improvement strategy. Those outcome measures still can be useful for informing quality improvement by identifying the need for stimulating new approaches to improvement.

Harmonization

Measures should be assessed for harmonization. Harmonization should be considered from the beginning of the development of the measure, and CMS contractors are expected to consider harmonization as one of the core measure evaluation criteria. Either the measure specifications must be harmonized with related measures so that they are uniform or compatible or the differences must be justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, risk adjustment, calculation, data source and collection instructions, and other measurement topics. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources. CMS contractors are expected to consider harmonization as one of the core measure evaluation criteria. (Refer to the Harmonization section for further details)

Related measures refer to measures that have either the same measure focus or target population. If similar or related measures were identified in the Information Gathering section, the measure contractor should document why the development of a new measure was deemed necessary, or why aspects of the measure were changed. Work closely with the Measures Manager to ensure that no duplication of measure development occurs. Harmonization should not result in inferior measures—measures should be based on the best measure concepts and ways to measure those concepts. There is no assumption that an endorsed measure is better than a new measure.

Measure contractors should also be aware that if the measure is to be submitted to NQF for endorsement, and there are other measures that essentially address the same target process, condition, event or outcome (numerator) and the same target population (denominator), the measures will be considered competing measures. The goal of NQF is to endorse the best measure and minimize confusing or conflicting information. Competing measures may already be endorsed or may be new submissions.

4.4 Applying the Measure Evaluation Criteria

In order to facilitate efficient and effective development of strong measures, the measures are evaluated throughout the development process by applying the standardized measure evaluation

criteria. The more consistent the measure properties are with the evaluation criteria the stronger the measure in terms of likelihood that it will be approved for use in a CMS program and endorsed by NQF. Through judicious application of the measure evaluation criteria at various times during measure development, measure developers should strive to identify weaknesses in the measure justification and technical specifications in order to revise and strengthen the measure, if possible. The measure developer should continuously update the Measure Information and Form (MIF) and Measure Justification with any information that can serve to demonstrate the strength of the measure.

The following key principles facilitate efficient and effective ongoing evaluation of the extent to which any measure is consistent with the measure evaluation criteria:

- "Importance to Measure and Report" serves as the primary threshold criterion. If the measure is to be submitted to NQF for endorsement, the measure must first pass all *Importance* subcriteria, or NQF will not evaluate it against the remaining criteria.
- "Scientific Acceptability" serves as the secondary threshold criterion. If the measure is to be submitted to NQF for endorsement, the measure must also pass all *Scientific Acceptability* subcriteria, or NQF will not evaluate it against the remaining criteria.
- The assessment of each criterion is a matter of degree. Not all acceptable measures will be strong—or equally strong—among each set of criteria.
- The measure evaluation process is iterative. Not all of the criteria can be evaluated in the early stages of measure development. Therefore, it will be necessary to leave some of the information on the measure evaluation report blank until a later time at which the information has been obtained to complete that aspect of the measure evaluation.
- The measure evaluation process is cumulative. The information obtained from each evaluation activity will be used to refine and strengthen the measure. Each successive measure evaluation report will reflect the results of the most recent evaluation activity.
- Harmonization is addressed. If there are related measures the evaluation process should compare the measures to address harmonization on an ongoing basis.

After all the criteria have been evaluated, the measure as a whole is evaluated, considering the summary ratings of each of the criteria.

The Measure Evaluation Guidance documents and Measure Evaluation report in the appendices facilitate a systematic approach for applying the measure evaluation criteria, rating the strength of the measure, and tracking the results to help the measure developer identify how to refine and strengthen the measure as it moves through the development and evaluation process. Although measure evaluation occurs throughout measure development, a formal measure evaluation report must be submitted to CMS to move the measure development process forward at two specific milestones:

1. When recommending approval of candidate measures for further development.
2. When recommending approval of fully tested and refined measures for implementation.

4.5 Using the Measure Evaluation Report and Measure Evaluation Guidance

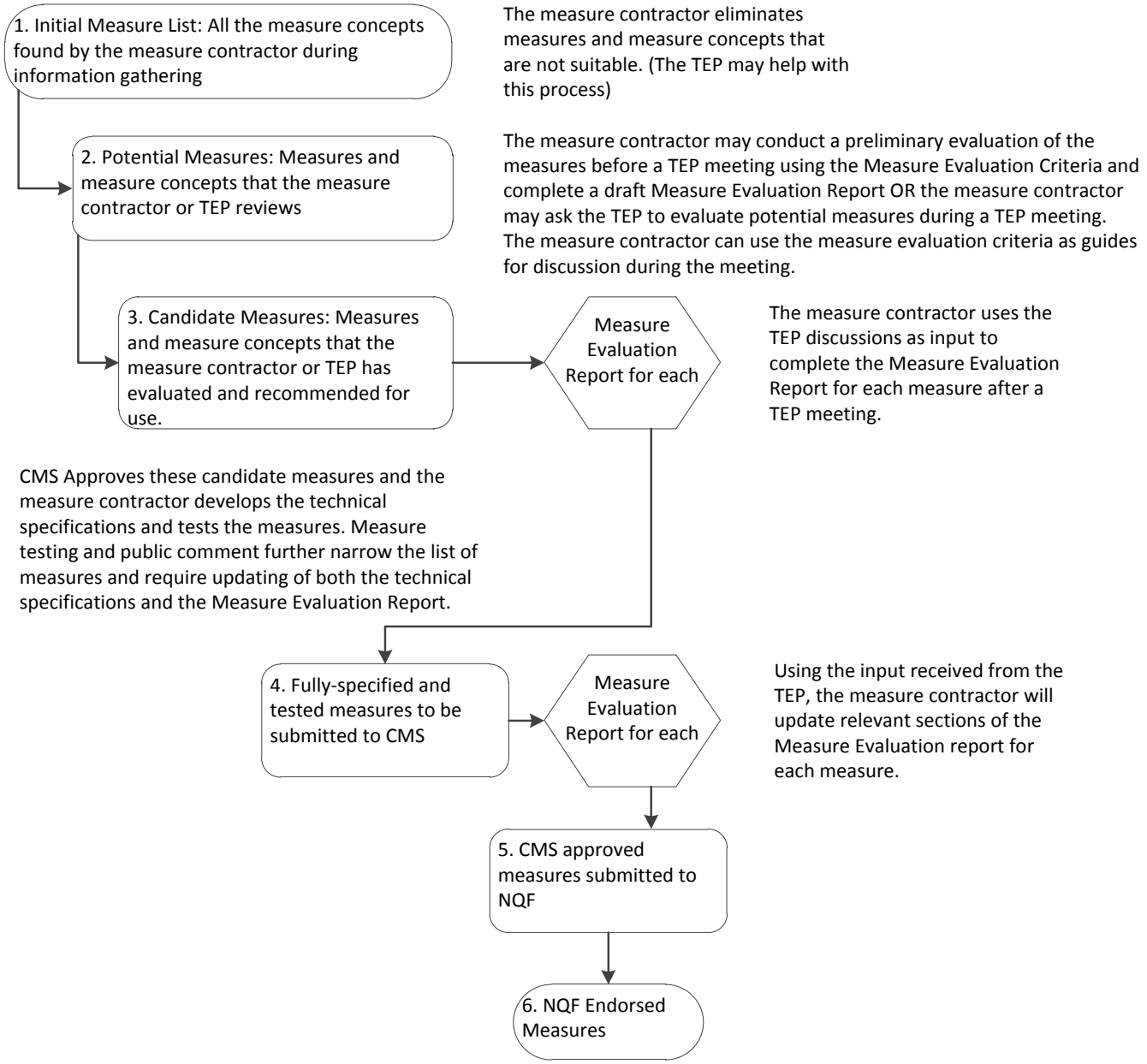
A measure evaluation report template and measure evaluation guidance documents corresponding to the measure evaluation subcriteria for individual measures, composite measures, and eMeasures are provided in the appendices in this section. Use of these materials is optional. They are designed to assist the contractor in applying the measure evaluation criteria and reporting the measure evaluation ratings in a formalized, standardized way. Information obtained by the measure contractor through the CMS Measures Management System's information gathering process, measure testing activities, TEP input, and from any public comment periods (as reflected in the MIF) can provide the basis for the measure contractor's application of the measure evaluation criteria, development of the measure evaluation report. It is important to evaluate the measure in an as objective manner as possible in order to anticipate any issues when the measure is submitted to NQF for endorsement. The measure evaluation report is where the contractor can communicate any anticipated risks associated with endorsement and present plans to strengthen any weaknesses identified. It is important for CMS to have an understanding of what it would take (pros/cons, costs/benefits) for increasing the rating and the risks if not undertaken.

Depending upon the particular needs of a contract, the Contracting Officer Representative/ Government Task Leader (COR/GTL) may instruct the measure contractor to use an alternative approach to measure evaluation. The Measure Evaluation Report can be modified as appropriate. The COR/GTL may direct that only certain criteria be evaluated or require measures to be evaluated more or less often during measure development.

4.6 Procedure

The measures are evaluated to determine the degree to which each measure is consistent with the standardized evaluation criteria. The resulting evaluation information is used to determine how the measure can be modified to increase the importance, scientific acceptability, usability and use, and feasibility of the measure. Figure 4-1 depicts the process of measure evaluation as a measure evolves from a conceptual level to a concrete level in preparation for implementation and illustrates possible instances for conducting a measure evaluation.

Figure 4-1 Measure Evaluation during Measure Development



4.7 Special Considerations

Certain types of measures require additional considerations when applying the Measure Evaluation Criteria. In addition to the discussion below, additional resources are available to assist measure contractors who may be evaluating these types of measures. These additional resources can be obtained from the Measures Manager and include customized Measure Evaluation Criteria for these specific measures types and Measure Evaluation Guidance that correspond to these customized sets of criteria.

Composite measures

A composite measure is a combination of two or more individual measures into a single measure that results in a single performance score. There are unique issues associated with composite methodology that require additional evaluation. The validity of the component measures; the appropriateness of the methods for scoring/aggregating and weighting the components; and interpretation of the composite score require evaluation. Both the composite and its component measures need to be evaluated to determine the suitability of the composite measure. The measure evaluation criteria and subcriteria include special considerations to be used when evaluating composite measures.

Below are principles to be used during the evaluation process. These eight principles for evaluating composite measures were adapted from the National Quality Forum (NQF) framework.[3]

- The components of the composite must be determined to meet the individual measure evaluation criteria as the first step in evaluating the composite measure. A component measure might not be important enough in its own right as an individual measure, but could be determined to be an important component of a composite.
- Even though all the component measures must individually meet the evaluation criteria, the composite measure as a whole also must meet the evaluation criteria.
- Composites may be developed beginning with a conceptual construct of quality or with a set of measures one wishes to summarize into one score. The methods used to develop and test the components must be justified.
- Methods for combining the component scores influence the interpretation of the composite measure results and must be justified.
- Rationale for choosing the analytic methods used to produce the quantifiable score needs to be justified.
- The methods for constructing a composite should be explicitly stated and transparent so that the composite can be deconstructed.
- The final composite result should be simple and readily interpretable by all stakeholders.
- The justification for the composite is the measure's effectiveness to accomplish the intended purpose for the composite measure (i.e., to assess, and ultimately improve, the quality of health care).

Additional evaluation subcriteria

In addition to the five measure evaluation criteria and related subcriteria, listed below are additional subcriteria that are used to evaluate composite measures:

[3]National Quality Forum. *Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety—Composite Measures.* August 2009. Available at: http://www.qualityforum.org/Projects/c-d/Composite_Evaluation_Framework/Composite_Evaluation_Framework_and_Composite_Measures.aspx. Accessed July 31, 2012.

Importance

- The component items/measures (e.g., types, focus) that are included in the composite are consistent with and representative of the conceptual construct for quality represented by the composite measure. Whether the composite measure development begins with a conceptual construct or a set of measures, the measures included must be conceptually coherent and consistent with the purpose.

Scientific Acceptability of the Measure Properties

- The composite measure is well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability. Composite specifications include methods for standardizing scales across component scores, scoring rules (i.e., how the component scores are combined or aggregated), weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite), handling of missing data, and required sample sizes.
- Component item/measure analysis (e.g., various correlation analyses such as internal consistency reliability), demonstrates that the included component items/measures fit the conceptual construct; or there must be justification and results for alternative analyses.
- Component item/measure analysis demonstrates that the included components contribute to the variation in the overall composite score; or if not, there is justification for why the component is included.
- The scoring/aggregation and weighting rules are consistent with the conceptual construct. Simple, equal weighting is often preferred unless differential weighting is justified. Differential weights are determined by empirical analyses or a systematic assessment of expert opinion or values-based priorities.
- Analysis of missing component scores supports the specifications for scoring/aggregation and handling of missing component scores.

Usability

- Data detail is maintained such that the composite measure can be decomposed into its components to facilitate transparency and understanding.

eMeasures

The electronic health record (EHR) holds significant promise for improving the measurement of health care quality. It can make available a broad range of reliable and valid data elements for quality measurement without the burden of data collection. Because clinical data are entered directly into standardized computer-readable fields, the EHR will be considered the authoritative source of clinical information and legal record of care.

For the most part, the evaluation criteria and subcriteria are the same for eMeasures as for measures developed using other data sources. However, evaluation of the *Scientific Acceptability, Validity,* and *Reliability* of eMeasures is based on some unique assumptions and special considerations, as follows:

- eMeasure evaluation is based on use of only data elements from the quality data model (QDM).
- Quality measures that are based on EHRs will automatically eliminate measurement errors due to manual abstraction, coding by persons other than the originator, or inaccurate transcription.
- eMeasures are subject to some of the same potential sources of error as non-eMeasures that could result in low evaluation ratings for the reliability and validity of data elements and measure scores including:
  - Incorrect measure specifications, including code lists, logic, or computer-readable programming language.
  - EHR system structure or programming that does not comply with standards for data fields, coding, or exporting data.
  - Data fields being used in different ways or entries made into the wrong EHR field.
- eMeasures are subject to an additional potential source of error from incorrect parsing of data by natural language processing software used to analyze information from text fields.
- Although data element reliability (repeatability) is assumed with computer programming of an EHR measure, empirical evidence is required to evaluate the reliability of the measure score.
- Data element validity for an eMeasure can be evaluated based on complete agreement between data elements and computed measure scores obtained by applying the eMeasure specifications to a simulated test EHR data set with known values for the critical data elements and computed measure score.
  - For retooled measures: Crosswalk of the eMeasure specifications (quality data model [QDM] elements, code lists, and measure logic) to the original measure specifications confirms that they represent the original measure, which was judged to be a valid indicator of quality.

Additional or adapted evaluation subcriteria

In addition to the standard five measure evaluation criteria and related subcriteria, listed below are additional or adapted subcriteria that are used to evaluate eMeasures.

Scientific Acceptability of the Measure Properties

- The measure is well-defined and precisely specified so it can be implemented consistently within and across organizations and allow for comparability, and EHR measure specifications are based on the quality data model (QDM).
- eMeasure specifications include data type from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting.
- Crosswalk of the EHR measure specifications (QDM quality data elements, code lists, and measure logic) to the endorsed measure specifications demonstrates that they represent the original measure, which was judged to be a valid indicator of quality.
- Data element: Validity demonstrated by analysis of agreement between data elements exported electronically and data elements abstracted from the entire EHR with statistical results within acceptable norms; OR, complete agreement between data elements and computed measure scores obtained by applying the EHR measure specifications to a simulated test EHR data set with known values for the critical data elements.

- Analysis of comparability of scores produced by the retooled EHR measure specifications with scores produced by the original measure specifications demonstrated similarity within tolerable error limits.

Cost and resource use measures

The resource use measure evaluation criteria are grounded in the standard National Quality Forum (NQF) evaluation criteria, keeping the five major criteria in place but modifying the subcriteria as appropriate to reflect the specific needs of resource use measure evaluation.

Resource use measures are broadly applicable and comparable measures of input counts (in terms of units or dollars) applied to a population or population sample. Resource use measures count the frequency of specific resources; these resource units may be monetized as appropriate. The approach to monetizing resources varies and often depends on the perspective of the measurer and those being measured. Monetizing resource use allows for the aggregation across resources.

Some general considerations related to evaluation of resource use measures include:

- Well-defined, complete, and precise specifications for resource use measures include: measure clinical logic and method, measure construction logic, and adjustments for comparability as relevant to the measure.
- Data protocol steps are critical to the reliability and validity of the measure.
- Examples of evidence that exclusion distorts measure results include, but are not limited to, frequency or cost of occurrence, sensitivity analyses with and without the exclusion, and variability of exclusions across providers.
- Some measures may specify the exclusion of some patients, events, or episodes that are known or determined to be high-cost. For example, a patient with active cancer may be excluded from a chronic obstructive pulmonary disease (COPD) resource use measure because cancer is considered to be the dominant medical condition with known high costs.
- Testing for resource use measure exclusions should address the appropriate specification steps (i.e., clinical logic, and thresholds and outliers).
- For those exclusions not addressed, justification for and implications of not addressing them is required.

4.8 Evaluation of Measure Set

Additional guiding principles are applied when selecting a set of measures for inclusion in a measure set. CMS developed standardized criteria for selecting measures for use across settings, programs and initiatives. The following core criteria and optional criteria may be applied when determining measures for inclusion in a set.

Core criteria

1. Measure addresses an important condition/topic with a performance gap and has a strong scientific evidence base to demonstrate that the measure when implemented can lead to the desired outcomes and/or more appropriate costs (i.e., NQF's Importance criteria)

2. Measure addresses one or more of the six National Quality Strategy Priorities (safer care, effective care coordination, preventing and treating leading causes of mortality and morbidity, person- and family-centered care, supporting better health in communities, making care more affordable)
3. Promotes alignment with specific program attributes and across CMS and the Department of Health and Human Services (HHS) programs
4. Program *measure set* includes consideration for health care disparities
5. Measure reporting is feasible

Optional criteria

6. Is responsive to specific program goals or statutory requirements
7. Enables measurement using measure type not already measured well (e.g., outcome, process, cost, etc.)
8. Enables measurement across the person-centered episode of care, demonstrated by assessment of the person's trajectory across providers and settings

4.9 Overview of Appendices

Below is a list of the appendices that are included in this section with brief explanation of their use:

Appendix 4a Tools for Evaluating Individual Measures

Individual Measure Evaluation Criteria and Subcriteria—This is the "basic" set of evaluation criteria and is appropriate for process, structure and outcome measures. It can be used to evaluate component measures within a composite.

Individual Measure Evaluation Guidance—This tool corresponds to the Individual Measure Evaluation Criteria and Subcriteria. Guidance is provided regarding how to assess/rate each of the criteria and subcriteria. The Measure Justification Form is aligned with the evaluation criteria and should include all the information necessary to evaluate the measure. The Guidance document should be used as the guide to determine the ratings for the criteria ratings that are reported in the Measure Evaluation report template, Appendix 4e.

Appendix 4b Tools for Evaluating Composite Measures

Composite Measure Evaluation Criteria and Subcriteria—This set of criteria contains additional criteria to evaluate composite measures. These criteria are based on criteria described in NQF's Composite Measure Evaluation Framework.

Composite Measure Evaluation Guidance—This tool corresponds to the Composite Measure Evaluation Criteria and Subcriteria. Guidance is provided regarding how to assess/rate each of the criteria and subcriteria. The Measure Justification Form is aligned with the evaluation criteria and should include all the information necessary to evaluate the measure. The Guidance document should be used as the guide to determine the ratings for the criteria ratings that are reported in the Measure Evaluation report template, Appendix 4e.

Appendix 4c Tools for Evaluating eMeasures

eMeasure Evaluation Criteria and Subcriteria—This set of criteria contains specific criteria to evaluate eMeasures. These criteria are based on criteria described in NQF's Measure Evaluation Criteria and Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties.

eMeasure Evaluation Guidance—This guidance tool corresponds to the eMeasure Evaluation Criteria and Subcriteria, and provides information to assess/rate each of the criteria and subcriteria. Use it as a guide in determining the criteria ratings that are reported in the Measure Evaluation Report (the template is located in Appendix 4e). The Measure Justification Form is aligned with the evaluation criteria and should include all the information necessary to evaluate the measure.

Appendix 4d Resource Use Measure Evaluation Criteria and Subcriteria

This set of criteria is based on NQF's Resource Use Measure Evaluation Criteria (Version 1.2).[4] At present, a guidance document has not been developed for cost and resource use measures; however, using the evaluation criteria, the Measure Evaluation report template can be used to document evaluation of the measure.

Appendix 4e Measure Evaluation Report Template

This form provides a template for documenting the formal review of each measure against the appropriate set of evaluation criteria. The instructions at the beginning of the document should be deleted when the document is used.

[4]National Quality Forum. *Resource Use Measure Evaluation Criteria (Version 1.2)*. November 2010. Available at: http://www.qualityforum.org/projects/efficiency_resource_use_1.aspx#t=2&s=&p=&e=1 Accessed: July 31, 2012.

Appendix 4a Tools for Evaluating Individual Measures

4a.1 Individual Measure Evaluation Criteria and Subcriteria

Adapted from National Quality Forum Measure Evaluation Criteria[5]

1. Impact, Opportunity, Evidence—Importance to Measure and Report:

Extent to which the specific measure focus is evidence-based, important to making significant gains in health care quality, and improving health outcomes for a specific high-impact aspect of health care where there is variation in or overall less-than-optimal performance.

1.a. High Impact

The measure focus addresses:

1.a.1 A specific national health goal/priority identified by Department of Health and Human Services (HHS) or the National Priorities Partnership convened by NQF,

OR

1.a.2 A demonstrated high-impact aspect of health care (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population, leading cause of morbidity/mortality, high resource use (current and/or future), severity of illness, and severity of patient/societal consequences of poor quality).

AND

1.b. Performance Gap

Demonstration of quality problems and opportunity for improvement (i.e., data demonstrating considerable variation, or overall less-than-optimal performance, in the quality of care across providers and/or population groups (disparities in care).

Note: Examples of data on opportunity for improvement include, but are not limited to, prior studies, epidemiologic data, or data from pilot testing or implementation of the proposed measure. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

AND

1.c. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

1.c.1 Health outcome: a rationale supports the relationship of the health outcome to processes or structures of care.

Note: Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

- *Intermediate clinical outcome, process, or structure*: A systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measure focus leads to a desired health outcome.

  Note: Clinical care processes typically include multiple steps: assess → identify problem/potential problem →

[5]National Quality Forum *Measure Evaluation* Criteria (January 2011). Available at: http://www.qualityforum.org/docs/measure_evaluation_criteria.aspx. Accessed July 31, 2012.

choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multi-step process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) **grading definitions** and **methods**, or Grading of Recommendations, Assessment, Development and Evaluation (**GRADE**) guidelines.

- *Patient experience with care*: Evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- *Efficiency*: Evidence for the quality component as noted above.

  Note: Measures of efficiency combine the concepts of resource use *and* quality (NQF's *Measurement Framework: Evaluating Efficiency Across Episodes of Care,* available at: http://www.qualityforum.org/Publications/2010/01/Measurement_Framework__Evaluating_Efficiency_Across_Patient-Focused_Episodes_of_Care.aspx, and *AQA Principles of Efficiency Measures* available at: http://www.aqaalliance.org/files/PrinciplesofEfficiencyMeasurement.pdf).

1.c.2 Measure focus is supported by the quantity of body of evidence, quality of body of evidence, and consistency of results of body of evidence.

- **Quantity of Body of Evidence:** Total number of studies (not articles or papers)
- **Quality of Body of Evidence:** Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence related to study factors including: study design or flaws; directness/indirectness to the specific measure (regarding the population, intervention, comparators, outcomes); imprecision (wide confidence intervals due to few patients or events). Study factors include: a) Study designs that affect certainty of confidence in estimates of effect include: randomized controlled trials (RCTs), which control for both observed and unobserved confounders, and non-RCTs (observational studies) with various levels of control for confounders, b) Study flaws that may bias estimates of effect include: lack of allocation concealment; lack of blinding; large losses to follow-up; failure to adhere to intention to treat analysis; stopping early for benefit; and failure to report important outcomes.
- **Consistency of Results of Body of Evidence:** Stability in both the direction and magnitude of clinically/practically meaningful benefits and harms to patients (benefit over harms) across studies in the body of evidence

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 1a, 1b, and 1c to pass this criterion or the NQF will not evaluate it against the remaining criteria.)

2. Reliability and Validity—Scientific Acceptability of Measure Properties:

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

2a. Reliability

2a1. The measure is well-defined and precisely specified so it can be implemented consistently within and across organizations and allow for comparability.

Note: Measure specifications include the target population (denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window, exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, and scoring/computation.

2a2. Reliability testing demonstrates that (1) the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period, and/or (2) that the measure score is precise.

Note: Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to, inter-rater/abstractor or intra-rater/abstractor studies, internal consistency for multi-item scales, and test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

2b. Validity

2b1. The measure specifications are consistent with the evidence presented to support the focus of measurement under criterion 1c. The measure is specified to capture the most inclusive target population indicated by the evidence, and exclusions are supported by the evidence.

2b2. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

Note: Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to, testing hypotheses that the measures scores indicate quality of care (e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method); correlation of measure scores with another valid indicator of quality for the specific topic; or, relationship to conceptually-related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion;

Note: Examples of evidence that an exclusion distorts measure results include, but are not limited to, frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

AND

If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

Note: Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- An evidence-based, risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; and has demonstrated adequate discrimination and calibration;

OR

- Rationale/data support no risk adjustment/ stratification.

Note: Risk factors that influence outcomes should not be specified as exclusions. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer

or inequalities in treatment for cardiovascular disease risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.
2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance, **OR** ◆ There is evidence of overall less-than-optimal performance. Note: With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.
2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.
2c. Disparities If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender); **OR** Rationale/data justifies why stratification is not necessary or not feasible.
(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria for both reliability and validity to pass this criterion, or the NQF will not evaluate it against the remaining criteria.)

3. Feasibility: Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement.
3a. For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).
3b. The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.
3c. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified.
3d. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). Note: All data collection must conform to laws regarding protected health information. Patient confidentiality is of

particular concern with measures based on patient surveys and when there are small numbers of patients.

4. Usability and Use: Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) are using or cold use performance results for both accountability and performance improvement to achieve the goal of high quality and efficient healthcare for individuals or populations.
4a. Accountability and Transparency Demonstration that performance results of a measure are used or can be used in public reporting, accreditation, licensure, health IT incentives, performance-based payment, or network inclusion/exclusion.
AND **4b. Improvement** Demonstration that performance results facilitate the goal of high quality efficient healthcare or credible rationale that the performance results can be used to further the goal of high quality efficient healthcare. Note: An important outcome that may not have an identified improvement strategy can still be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.
AND **4c.** Benefits of the performance measure in facilitating progress toward achieving high quality efficient healthcare outweigh the evidence of unintended consequences to individuals or populations (if such evidence exists).

5. Harmonization: Extent to which either measure specifications are harmonized with related measures so they are uniform or compatible or the differences must be justified (e.g. dictated by evidence).
5a. Related measure: The measure specifications for this measure are completely harmonized with a related measure.
5b. Competing measure: This measure is superior to competing measures (e.g. a more valid or efficient way to measure quality); OR has additive value as an endorsed additional measure (provide analyses if possible).

4a.2 Individual Measure Evaluation Guidance

(Evaluation criteria adapted from National Quality Forum (NQF) evaluation criteria)

Measure Name:

Measure Set:

Type of Measure:

Instructions: For each subcriterion, check the description that best matches your assessment of the measure. Use the supporting information provided in the Measure Information Form (MIF) and Measure Justification, as well as any additional relevant studies or data. Based on your rating of the subcriteria, use the Measure Evaluation Criteria Summary Rating guidelines included in this tool to make a summary determination for each criterion. Use the information to complete the Measure Evaluation report template.

Importance—Impact, Opportunity, Evidence

Subcriterion	Pass	Fail
1a. High Impact	The measure focus addresses a specific national health goal/priority identified by one or more of the following: ◆ CMS/HHS ◆ Legislative mandate ◆ NQF's National Priorities Partners **OR** The measure focus has high impact on health care as demonstrated by one or more of the following: ◆ Affects large numbers ◆ Substantial impact for a small population ◆ A leading cause of morbidity/mortality ◆ Severity of illness ◆ High Resource Use ◆ Potential cost savings to the Medicare Program (business case[6]) ◆ Patient/societal consequences of poor quality regardless of cost (social case)	The measure does not directly address a national health goal/priority. **AND** The data do not indicate it is a high impact aspect of health care, or is unknown.

[6]A business case is described in Section 5, Information Gathering - Appendix 5-C of the Blueprint.

1b. Performance Gap	Evidence exists to substantiate a quality problem and opportunity for improvement (i.e., data demonstrate considerable variation) **OR** Data demonstrate overall poor performance across providers or population groups (disparities).	Performance gap is unknown, **OR** There is limited or no room for improvement (no variability across providers or population groups and overall good performance).

Subcriterion	High	Moderate	Low
1c: **Quantity of body of evidence**: Total number of studies (not articles or papers)	5+ studies	2-4 studies	0-1 studies
1c: **Quality of body of evidence**	Randomized controlled trials (RCTs) of direct evidence, with adequate size to obtain precise estimates of effect, and without serious flaws that introduce bias.	Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect; **OR** RCTs without serious flaws that introduce bias, but with either indirect evidence, or imprecise estimate of effect.	RCTs with flaws that introduce bias **OR** Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations
1c: **Consistency of body of evidence**	Estimates of benefits and harms to patients are consistent in direction and similar in magnitude across studies in the body of evidence.	Estimates of benefits and harms to patients are consistent in direction, but differ in magnitude across studies in the body of evidence; **OR** If only one study, the estimate of benefits greatly outweighs the estimate of potential harms, **OR** For expert opinion that is systematically assessed,	Differences in both magnitude and direction of benefits and harms to patients across studies in the body of evidence, or wide confidence intervals prevent estimating net benefit; **OR** For expert opinion evidence that is systematically assessed, lack of agreement that benefits to patients clearly outweigh potential harms.

		agreement that benefits to patients clearly outweigh potential harms.	
1c: **Potential Exception to Empirical Body of Evidence – (**structure and process measures)	*High does not apply. If this exception is applicable, 1c is either rated Moderate or Low*	If there is no empirical evidence, expert opinion is systematically assessed with agreement that the benefits to patients greatly outweigh potential harms.	For expert opinion evidence that is systematically assessed, lack of agreement that benefits to patients clearly outweigh potential harms.
1c: **Exception to Empirical Body of Evidence** – (health outcome measures)	Empirical evidence links the health outcome to a known process or structure of care.	A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service.	No rationale is given that supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service.

Guidelines for Summary Rating: Importance

Instructions for evaluating subcriterion 1c Body of Evidence: In order to determine if the measure passes subcriterion 1c body of evidence, use the applicable table below. After evaluating 1c, determine if measure meets Importance by having passed all of the subcriteria (1a, 1b, and 1c).

Structure and Process Measures – rating of body of evidence (1c)

Quantity of Body of Evidence	Quality of Body of Evidence	Consistency of Results of Body of Evidence	Pass Subcriterion 1c
Moderate-High	Moderate-High	Moderate-High	Yes
Low	Moderate-High	Moderate (if only one study, high consistency not possible)	Yes, but only if it is judged that additional research is unlikely to change conclusion that benefits to patients outweigh harms; otherwise, No
Moderate-High	Low	Moderate-High	Yes, but only if it is judged that potential benefits to

			patients clearly outweigh potential harms; otherwise, No
Low-Moderate-High	Low-Moderate-High	Low	No
Low	Low	Low	No
No empirical evidence ***(potential exception)***	No empirical evidence ***(potential exception)***	No empirical evidence ***(potential exception)***	Yes, but only if it is systematically judged by an expert panel that potential benefits to patients clearly outweigh potential harms; otherwise, No

Health Outcome Measures – exception to rating body of evidence (1c)

Process, Structure or Intervention Linkage to Outcome	Pass Subcriterion 1c
High-Moderate	Yes
Low	No
Summary Rating: Importance **Pass**: A**ll** of the subcriteria (1a, 1b, 1c) are rated "**Pass".** **Fail**: A**ny** of subcriteria (1a, 1b, 1c) are rated **"Fail".** *(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 1a, 1b, and 1c to pass this criterion, or NQF will not evaluate it against the remaining criteria.)*	

Scientific Acceptability of Measure Properties —Reliability, Validity, Disparities

2a. Reliability:

Instructions: To rate "High" for reliability; both subcriteria need to have a "High" rating. If there is a combination of "high" and moderate" the overall Reliability rating is "Moderate". If the measure meets any of the definitions in "Low" column, Reliability will be rated "Low" and the measure will fail.

Subcriterion	High	Moderate	Low
2a1. Specifications well defined and precisely specified	All measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring, etc.) are unambiguous and likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.;	*Moderate does not apply. This subcriterion is either rated High or Low.*	One or more measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) are ambiguous with potential for confusion in identifying who is included and excluded from the target population, or the event, condition, or outcome being measured; or how to compute the score, etc.;
2a2. Reliability testing	Empirical evidence of reliability of *BOTH data elements **AND** measure score within acceptable norms*. **Data element**: appropriate method, scope, and reliability statistics for critical data elements within acceptable norms (new testing, or prior evidence for the same data type); **OR** commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission); **OR** *may forego data element reliability testing if data element validity was demonstrated*; **AND** **Measure score**: appropriate method, scope, and reliability statistics for score computation or	Empirical evidence of reliability *within acceptable norms for either critical data elements **OR** measure score*. **Data element**: appropriate method, scope, and reliability statistics for critical data elements within acceptable norms (new testing, or prior evidence for the same data type); **OR** commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission); **OR** *may forego data element reliability testing if data element validity was demonstrated*; **OR** **Measure score**: appropriate method, scope, and reliability statistics for score computation or	Empirical evidence (using appropriate method and scope) of *unreliability for either data elements **OR** measure score,* i.e., statistical results outside of acceptable norms **OR** Inappropriate method or scope of reliability testing

Subcriterion	High	Moderate	Low
	risk adjustment	risk adjustment	

2b. Validity

Instructions: To rate "High" for Validity; all subcriteria must have a "High" rating. If there is a combination of "High" and "Moderate" the overall Validity rating is "Moderate". If the measure meets any of the definitions in "Low" column, Validity will be rated "Low" and the measure will fail.

Subcriterion	High	Moderate	Low
2b1. Measure specifications	The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1c) under *Importance to Measure and Report*	*Moderate does not apply. This subcriterion is either rated High or Low*	The measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above;
2b2. Validity testing	Empirical evidence of validity of ***BOTH*** *data elements* ***AND*** *measure score within acceptable norms*: **Data element**: appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; **Measure score**: Evidence that supports the intended interpretation of measure scores for the intended purpose—making conclusions about the quality of care. Examples of the types of measure score validity testing: • Construct validity • Discriminative validity/Contrasted groups • Predictive validity	Empirical evidence of validity *within acceptable norms for either critical data elements* ***OR*** *measure score* **Data element**: appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; **Measure score**: Evidence that supports the intended interpretation of measure scores for the intended purpose—making conclusions about the quality of care. Examples of the types of measure score validity testing: • Construct validity • Discriminative validity/Contrasted groups • Predictive validity	Empirical evidence (using appropriate method and scope) of *invalidity for either data elements* ***OR*** *measure score*, i.e., statistical results outside of acceptable norms **OR** Systematic assessment of face validity of measure resulted in *lack of consensus* as to whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and poor quality. **OR** Inappropriate method or scope of validity testing (including inadequate assessment of face validity)

Subcriterion	High	Moderate	Low
	• Convergent validity • Reference strategy/Criterion validity	• Convergent validity Reference strategy/Criterion validity **OR** Systematic assessment of face validity of measure, which is the extent to which a measure appears to reflect that which it is supposed to measure "at face value." Face validity for a CMS quality measure may be adequate if accomplished through a systematic and transparent process, by a panel of experts, such as the TEP, where formal rating of the validity is recorded and appropriately aggregated. The TEP should explicitly address whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and poor quality.	
2b2. Validity testing – threats to validity	Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased	*Moderate does not apply. This subcriterion is either rated High or Low*	Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and **determined to bias results** **OR** Threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) **are likely and are NOT empirically assessed**
2b3. Exceptions	Exceptions are supported by the clinical evidence, otherwise they are supported by evidence of sufficient frequency of occurrence	*Moderate does not apply. This subcriterion is either rated High or Low*	Exceptions are not supported by evidence, **1. OR**

Subcriterion	High	Moderate	Low
	so that results are distorted without the exclusion; **AND** Measure specifications for scoring include computing exceptions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); **AND** If patient preference (e.g., informed decision-making) is a basis for exception, there must be evidence that the exception impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exception category computed separately).		The effects of the exceptions are not transparent. (i.e., impact is not clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); **OR** If patient preference (e.g., informed decision-making) is a basis for exception, there is no evidence that the exception impacts performance on the measure;
2b4. Risk Adjustment ***For outcome measures and other measures (e.g., resource use) when indicated***:	An evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are present at start of care, **AND** uses scientifically sound methods; **OR** rationale/data support not using risk adjustment.	An evidence-based risk-adjustment strategy is specified consistent with the evaluation criteria, HOWEVER it uses a method that is less than ideal, but acceptable.	A risk-adjustment strategy is not specified AND would be absolutely necessary for the measure to be fair and support valid conclusions about the quality of care.
2b5.	Data analysis demonstrates that methods for scoring and analysis	Data analysis was conducted but did not demonstrate statistically	Data analysis was not conducted to identify statistically significant

Subcriterion	High	Moderate	Low
Meaningful	allow for identification of statistically significant and practically/clinically meaningful differences in performance. **OR** there is evidence of overall less than optimal performance.	significant and practically/clinically meaningful differences in performance; HOWEVER the methods for scoring and analysis are appropriate to identify such differences if they exist.	and practically/clinically meaningful differences in performance, **OR** Methods for scoring and analysis are not appropriate to identify such difference.
2b6. Comparable results	***If multiple data sources/methods are allowed*** (specified in the measure), data and analysis demonstrate that they produce comparable results	Multiple data sources/methods are allowed with no formal testing to demonstrate they produce comparable results, HOWEVER there is a credible description of why/how the data elements are equivalent and the results should be comparable	Multiple data sources/methods are allowed and it is unknown if they produce comparable results, **OR** testing demonstrates they do not produce comparable results
2c. Disparities			
	High	**Moderate**	**Low**
	Data indicate disparities do not exist; 2. **OR** if disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results	Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities; HOWEVER the rationale/data justify why stratification is neither necessary nor feasible	Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities; **AND** rationale/data are not provided to justify why stratification is neither necessary nor feasible

Guidelines for Summary Rating: Scientific Acceptability of Measure Properties

Instructions: If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 2a, 2b, and 2c to pass this criterion, or NQF will not evaluate it against the remaining criteria. Reliability and validity (2a and 2b) are assessed in combination. In order to determine if the measure passes reliability and validity, use the table below.

Validity Rating	Reliability Rating	Pass	Description
High	Moderate-High	Yes	Evidence of reliability and validity
High	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Moderate	Moderate-High	Yes	Evidence of reliability and validity
Moderate	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Low	Any rating	No	Validity of conclusions about quality is the primary concern. If evidence of validity is rated low, the reliability rating will usually also be low. Low validity and moderate-high reliability represents inconsistent evidence.

Summary Rating: Scientific Acceptability of Measure Properties

Pass: The measure rates **moderate to high on all** aspects of reliability **and** validity **and** disparities

Fail: The measure **rates low** for one or more aspects of reliability **or** validity or disparities

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria for both reliability and validity to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

3. Usability

Instructions: If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass both subcriteria 3a and 3b. A measure must be rated High or Moderate both public reporting and quality improvement usability to pass

Outcome Measures: An important **outcome** that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement and can be rated “High” or “Moderate”.

Subcriterion	High	Moderate	Low
3a. Public Reporting	Testing demonstrates that information produced by the measure is meaningful, understandable, and useful for **public reporting** (e.g., systematic feedback from users, focus group, cognitive testing).	Formal testing has not been performed, but the measure is in widespread use and you think it is meaningful and understandable for **public reporting** (e.g., focus group, cognitive testing) **OR** *When measure is being rated during its initial development*: A rationale for how the measure performance results will be meaningful, understandable, and useful for public reporting.	The measure is not in use and has not been tested for usability; **OR** Testing demonstrates information produced by the measure is **not meaningful**, understandable, and useful for public reporting **OR** *When measure is being rated during its initial development*: A **rationale** for how the measure performance results will be meaningful, understandable, and useful for public reporting is not provided.
3b. Quality Improvement	Testing demonstrates that information produced by the measure is meaningful, understandable, and useful for **quality improvement** (e.g., systematic feedback from users, analysis of quality improvement initiatives).	Formal testing has not been performed but the measure is in widespread use and accepted to be meaningful and useful for **quality improvement** (e.g., quality improvement initiatives). **OR** *When measure is being rated during its initial development*: A rationale for how the measure performance results will be meaningful, understandable, and useful for quality improvement.	The measure is not in use and has not been tested for usability; **OR** Testing demonstrates information produced by the measure is **not meaningful**, understandable, and useful for public reporting **OR** *When measure is being rated during its initial development*: A **rational**e for how the measure performance results will be meaningful, understandable, and useful for quality improvement is **not provided**.

Guidelines for Summary Rating: Usability

Summary Rating: Usability

Pass: The measure rates "M**oderate"** to **"High"** on **both** aspects usability

Fail: The measure rates **"Low"** for **one or both** aspects of usability

4. **Feasibility**.

Subcriterion	High	Moderate	Low
4a. Byproduct of care ***(clinical measures only)***	The required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery (e.g., BP reading, diagnosis).	The required data are based on information generated during care delivery; HOWEVER, trained coders or abstractors are required to use the data in computing the measure.	The required data are not generated during care delivery and are difficult to collect or require special surveys or protocols.
4b. Electronic data	The required data elements are available in electronic health records or other electronic sources.	If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.	The required data elements are not available in electronic sources **AND** There is no credible, near-term path to electronic collection specified.
4c. Inaccuracies, errors, or unintended consequences	Susceptibility to inaccuracies, errors, or unintended consequences related to measurement are judged to be inconsequential or can be minimized through proper actions **OR** can be monitored and detected	There is moderate susceptibility to inaccuracies, errors, or unintended consequences, **AND/OR** They are more difficult to detect through auditing.	Inaccuracies, errors, or unintended consequences have been demonstrated, **AND** They are not easily detected through auditing.
4d. Data collection strategy	The measure is in operational use and the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) has been implemented without difficulty.	The measure is not in operational use; **HOWEVER** testing demonstrates the data collection strategy can be implemented with minimal difficulty or additional resources.	The measure is not in operational use, **AND** Testing indicates the data collection strategy was difficult to implement and/or requires

substantial additional resources.

Guidelines for Summary Rating: Feasibility

Summary Rating: Feasibility

High rating indicates: **Three or four** subcriteria are rated "**High**".

Moderate rating indicates: "**Moderate" or mixed ratings**, with **no more than one "Low"** rating.

Low rating indicates: **Two or more** subcriteria are rated **"Low".**

Harmonization

Instructions: Measures should be assessed for harmonization. Either the measure specifications must be harmonized with related measures so that they are uniform or compatible or the differences must be justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources. Harmonization should not result in inferior measures—measures should be based on the best measure concepts and ways to measure those concepts. There is no presupposition that an endorsed measure is better than a new measure.

Completely	Partially	Not Harmonized
The measure specifications are **completely harmonized** with related measures; the measure can be used at multiple levels or settings/data sources	The measure specifications are **partially harmonized** with related measures, HOWEVER the rationale justifies any differences; the measure can be used at one level or setting/data source	The measure specifications are **not harmonized** with related measures **AND** the rationale does not justify the differences

5. Harmonization

Instructions: Measures should be assessed for harmonization. Either the measure specifications must be harmonized with related measures so that they are uniform or compatible or the differences must be justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources. Harmonization should not result in inferior measures—measures should be based on the best measure concepts and ways to measure those concepts. There is no presupposition that an endorsed measure is better than a new measure.

Completely	Partially	Not Harmonized
The measure specifications are **completely harmonized** with related measures; the measure can be used at	The measure specifications are **partially harmonized** with related measures, HOWEVER the rationale	The measure specifications are **not harmonized** with related measures

Completely	Partially	Not Harmonized
multiple levels or settings/data sources. AND There are no competing measures (already endorsed, in development, or in use)	justifies any differences; the measure can be used at one level or setting/data source	**AND** the rationale does not justify the differences OR There is a competing measure (same focus, same population)

Appendix 4b Tools for Evaluating Composite Measures

4b.1 Composite Measure Evaluation Criteria and Subcriteria

Adapted from National Quality Forum Measure Evaluation Criteria[7] and Composite Measure Evaluation Framework[8]

1. Impact, Opportunity, Evidence—Importance to Measure and Report:

Extent to which the specific measure focus is evidence-based, important to making significant gains in health care quality, and improving health outcomes for a specific high-impact aspect of health care where there is variation in or overall less-than-optimal performance.

If the component measures are determined to meet subcriteria 1a, 1b, and 1c, then the composite would meet 1a, 1b, and 1c. If a component measure does not meet 1a, 1b, and 1c and is not important enough in its own right as an individual measure, it could still meet the Importance to Measure and Report criterion if it is determined to be an important component of a composite and meets subcriteria 1d and 1e.

1a. High Impact

For each component measure of the composite measure, the measure focus addresses:

- A specific national health goal/priority identified by the Department of Health and Human Services (HHS) or the National Priorities Partnership convened by NQF;

OR

- A demonstrated high-impact aspect of health care (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population, leading cause of morbidity/mortality, high resource use (current and/or future), severity of illness, and severity of patient/societal consequences of poor quality).

AND

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement (i.e., data demonstrating considerable variation, or overall less-than-optimal performance, in the quality of care across providers and/or population groups (disparities in care)).

Note: Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, or data from pilot testing or implementation of the proposed measure. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

AND

1c. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

Health outcome: A rationale supports the relationship of the health outcome to processes or structures of care.

[7] National Quality Forum *Measure Evaluation* Criteria (January 2011). Available at: http://www.qualityforum.org/docs/measure_evaluation_criteria.aspx. Accessed July 31, 2012.

[8] National Quality Forum *Composite Evaluation Framework* (August 2009). Available at: http://www.qualityforum.org/Projects/c-d/Composite_Evaluation_Framework/Composite_Evaluation_Framework_and_Composite_Measures.aspx Last Accessed June 2011.

Note: Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

Intermediate clinical outcome, **process**, or **structure**: A systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measure focus leads to a desired health outcome.

Note: Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) **grading definitions** and **methods**, or Grading of Recommendations, Assessment, Development and Evaluation (**GRADE**) **guidelines**.

Patient experience with care: Evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
Efficiency: Evidence for the quality component as noted above.

Note: Measures of efficiency combine the concepts of resource use **and** quality (NQF's *Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures*).

Measure focus is supported by the quantity of body of evidence, quality of body of evidence, and consistency of results of body of evidence.

- **Quantity of Body of Evidence:** Total number of studies (not articles or papers)
- **Quality of Body of Evidence:** Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence related to study factors including: study design or flaws; directness/indirectness to the specific measure (regarding the population, intervention, comparators, outcomes); imprecision (wide confidence intervals due to few patients or events). Study factors include: a) Study designs that affect certainty of confidence in estimates of effect include: randomized controlled trials (RCTs), which control for both observed and unobserved confounders, and non-RCTs (observational studies) with various levels of control for confounders, b) Study flaws that may bias estimates of effect include: lack of allocation concealment; lack of blinding; large losses to follow-up; failure to adhere to intention to treat analysis; stopping early for benefit; and failure to report important outcomes.

Consistency of Results of Body of Evidence: Stability in both the direction and magnitude of clinically/practically meaningful benefits and harms to patients (benefit over harms) across studies in the body of evidence

1d. The purpose/objective of the composite measure and the construct for quality are clearly described.

1e. The component items/measures (e.g., types, focus) that are included in the composite are consistent with and representative of the conceptual construct for quality represented by the composite measure. Whether the composite measure development begins with a conceptual construct or a set of measures, the measures included must be conceptually coherent and consistent with the purpose.

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 1a, 1b, 1c, 1d and 1e to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

2. Reliability and Validity—Scientific Acceptability of Measure Properties:
Extent to which the measure**, as specified**, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.
2a. Reliability **2a1.** The measure is well-defined and precisely specified so it can be implemented consistently within and across organizations and allow for comparability and electronic health record (EHR) measure specifications are based on the quality data model (QDM). Note: **Individual component measure specifications** include the target population (denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window, exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, scoring/computation. **Composite specifications** include methods for standardizing scales across component scores, scoring rules (i.e., how the component scores are combined or aggregated), weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite), handling of missing data, and required sample sizes.
2a2. Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. Note: eMeasures specifications include data types from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to inter-rater/abstractor or intra-rater/abstractor studies, internal consistency for multi-item scales, and test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
2b. Validity **2b1.** The measure specifications are consistent with the evidence presented to support the focus of measurement under criterion 1c. The measure is specified to capture the most inclusive target population indicated by the evidence, and exclusions are supported by the evidence.
2b2. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. If face validity is the only validity addressed, it is systematically assessed. Note: Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: Testing hypotheses that the measure scores indicate quality of care (e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually-related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality
2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency

of occurrence so that results are distorted without the exclusion;

Note: Examples of evidence that an exclusion distorts measure results include, but are not limited to frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

AND

If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

Note: Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- An evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; and has demonstrated adequate discrimination and calibration;

OR

- Rationale/data support no risk adjustment/ stratification.

Note: Risk factors that influence outcomes should not be specified as exclusions. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified composite measure allow for identification of statistically significant and practically/clinically meaningful differences in performance,

OR

There is evidence of overall less-than-optimal performance.

Note: With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. Component item/measure analysis (e.g., various correlation analyses such as internal consistency reliability), demonstrates that the included component items/measures fit the conceptual construct;

OR

Justification and results for alternative analyses are provided.

2b8. Component item/measure analysis demonstrates that the included components contribute to the variation in the overall composite score;

OR

If not, justification for inclusion is provided.

2b9. The scoring/aggregation and weighting rules are consistent with the conceptual construct. (Simple, equal weighting is often preferred unless differential weighting is justified. Differential weights are determined by empirical analyses or a systematic assessment of expert opinion or values-based priorities.)

2b10. Analysis of missing component scores supports the specifications for scoring/aggregation and handling of missing component scores.

2c. Disparities

If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender);

OR

Rationale/data justifies why stratification is not necessary or not feasible.

(Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.) If measure does not meet all subcriteria for reliability and validity, STOP; the evaluation does not proceed.

3. Feasibility:
Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. For clinical composite measures, overall the required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3b. The required data elements for the composite overall are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3c. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified.

3d. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) for obtaining all component measures can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

Note: All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.

4. Usability and Use:
Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) are using or cold use performance results for both accountability and performance improvement to achieve the goal of high quality and efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Demonstration that performance results of a measure are used or can be used in public reporting, accreditation, licensure, health IT incentives, performance-based payment, or network inclusion/exclusion.

AND **4b. Improvement** Demonstration that performance results facilitate the goal of high quality efficient healthcare or credible rationale that the performance results can be used to further the goal of high quality efficient healthcare. Note: An important outcome that may not have an identified improvement strategy can still be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.
4c. Review of existing endorsed measures and measure sets demonstrates that the composite measure provides a distinctive or additive value to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of health care, is a more valid or efficient way to measure).
4d. Data detail is maintained such that the composite measure can be decomposed into its components to facilitate transparency and understanding.
4e. Demonstration (through pilot testing or operational data) that the composite measure achieves the stated purpose/objective.
5. Harmonization: Extent to which either measure specifications are harmonized with related measures so they are uniform or compatible, or the differences must be justified (e.g. dictated by evidence).
5a. Related measure: The measure specifications for this measure are completely harmonized with a related measure.
5b. Competing measure: This measure is superior to competing measures (e.g. a more valid or efficient way to measure quality); OR has additive value as an endorsed additional measure (provide analyses if possible)

4b.2 Composite Measure Evaluation Guidance

(Evaluation criteria adapted from National Quality Forum (NQF) evaluation criteria)

Instructions: For each subcriterion, check the description that best matches your assessment of the measure. Consider both the composite measure and each component separately. Use the supporting information provided in the Measure Information Form (MIF) and Measure Justification, as well as any additional relevant studies or data. Based on your rating of the subcriteria, use the Measure Evaluation Criteria Summary Rating guidelines included in this tool to make a summary determination for each criterion. Use the information to complete the Measure Evaluation report template.

Importance—Impact, Opportunity, Evidence

Subcriterion	Pass	Fail
1a. High Impact	The measure focus addresses a specific national health goal/priority identified by one or more of the following: CMS/HHS Legislative mandate NQF's National Priorities Partners **OR** The measure focus has high impact on health care as demonstrated by one or more of the following: Affects large numbers Substantial impact for a small population A leading cause of morbidity/mortality Severity of illness High Resource Use Potential cost savings to the Medicare Program (business case[9]) Patient/societal consequences of poor quality	The measure does not directly address a national health goal/priority. **AND** The data do not indicate it is a high impact aspect of health care, or is unknown.

[9] A business case is described in Section 5, Information Gathering - Appendix 5-C of the Blueprint.

Subcriterion	Pass		Fail
	regardless of cost (social case)		
1b. Performance Gap	Evidence exists to substantiate a quality problem and opportunity for improvement (i.e., data demonstrate considerable variation) **OR** Data demonstrate overall poor performance across providers or population groups (disparities).		Performance gap is unknown, **OR** There is limited or no room for improvement (no variability across providers or population groups and overall good performance).
1c: **Quantity of body of evidence**: Total number of studies (not articles or papers)	5+ studies	2-4 studies	0-1 studies
1c: **Quality of body of evidence**	Randomized controlled trials (RCTs) of direct evidence, with adequate size to obtain precise estimates of effect, and without serious flaws that introduce bias.	Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect; **OR** RCTs without serious flaws that introduce bias, but with either indirect evidence, or imprecise estimate of effect.	RCTs with flaws that introduce bias **OR** Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations
1c: **Consistency of body of evidence**	Estimates of benefits and harms to patients are consistent in direction and similar in magnitude across studies in the body of evidence.	Estimates of benefits and harms to patients are consistent in direction, but differ in magnitude across studies in the body of evidence; **OR** If only one study, the estimate of benefits greatly outweighs	Differences in both magnitude and direction of benefits and harms to patients across studies in the body of evidence, or wide confidence intervals prevent estimating net benefit; **OR** For expert opinion evidence that is systematically assessed, lack of agreement that benefits to patients

Subcriterion	Pass	Fail	
		the estimate of potential harms, **OR** For expert opinion that is systematically assessed, agreement that benefits to patients clearly outweigh potential harms.	clearly outweigh potential harms.
1c: **Potential Exception to Empirical Body of Evidence –** (structure and process measures)	*High does not apply. If this exception is applicable, 1c is either rated Moderate or Low*	If there is no empirical evidence, expert opinion is systematically assessed with agreement that the benefits to patients greatly outweigh potential harms.	For expert opinion evidence that is systematically assessed, lack of agreement that benefits to patients clearly outweigh potential harms.
1c: **Exception to Empirical Body of Evidence** – (health outcome measures)	Empirical evidence links the health outcome to a known process or structure of care.	A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service.	No rationale is given that supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service.

Instructions for evaluating subcriterion 1c Body of Evidence: In order to determine if the measure passes subcriterion 1c body of evidence, use the applicable table below. After evaluating 1c, determine if measure meets Importance by having passed all of the subcriteria (1a, 1b, and 1c).

Structure and Process Measures –rating body of evidence (1c)

Quantity of Body of Evidence	Quality of Body of Evidence	Consistency of Results of Body of Evidence	Pass Subcriterion 1c
Moderate-High	Moderate-High	Moderate-High	Yes

Low	Moderate-High	Moderate (if only one study, high consistency not possible)	Yes, but only if it is judged that additional research is unlikely to change conclusion that benefits to patients outweigh harms; otherwise, No
Moderate-High	Low	Moderate-High	Yes, but only if it is judged that potential benefits to patients clearly outweigh potential harms; otherwise, No
Low-Moderate-High	Low-Moderate-High	Low	No
Low	Low	Low	No
No empirical evidence ***(potential exception)***	No empirical evidence ***(potential exception)***	No empirical evidence ***(potential exception)***	Yes, but only if it is systematically judged by an expert panel that potential benefits to patients clearly outweigh potential harms; otherwise, No

Health Outcome Measures – exception to rating body of evidence (1c)

Process, Structure or Intervention Linkage to Outcome	Pass Subcriterion 1c
High-Moderate	Yes
Low	No

Composite Measure Specific Criteria:

Subcriterion	Pass	Fail
1a, 1b, and 1c Component ratings	The component measures are determined to meet the importance criteria 1a, 1b, and 1c, **OR** A component measure is not important enough in its own right as an individual measure, but is an	a component measure is not important enough in its own right as an individual measure, **And** is not an important component of the composite

	important component of the composite.	
1d Purpose/objective	The purpose/objective of the composite measure and the construct for quality is described	The purpose/objective of the composite measure and the construct for quality is not described
1e Components consistent with construct	The component items/measures (e.g., types, focus) that are included in the composite are consistent with and representative of the conceptual construct for quality represented by the composite measure.	The component items/measures that are included in the composite are not consistent with and representative of the conceptual construct for quality represented by the composite measure.

Guidelines for Summary Rating: Importance

Summary Rating: Importance

Pass: A**ll** of the subcriteria (1a, 1b, 1c, 1d, and 1e) are rated "**Pass".**

Fail: A**ny** of subcriteria (1a, 1b, 1c, 1d, and 1e) are rated **"Fail".**

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 1a, 1b, 1c, 1d, and 1e to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

Scientific Acceptability of Measure Properties —Reliability, Validity, Disparities

2a. Reliability:

Instructions: To rate "High" for reliability; both subcriteria need to have a "High" rating. If there is a combination of "high" and moderate" the overall Reliability rating is "Moderate". If the measure meets any of the definitions in "Low" column, Reliability will be rated "Low" and the measure will fail.

Subcriterion	High	Moderate	Low
2a1. Specifications well defined and precisely	All measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring, etc.) are unambiguous and likely to consistently identify who is	All measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring, etc.) are unambiguous and likely to consistently identify who is	One or more measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) are ambiguous with potential for confusion in

Subcriterion	High	Moderate	Low
specified	included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.; **AND** Composite specifications include **all of the following**: methods for standardizing scales across component scores, scoring rules, weighting rules, handling of missing data, and required sample sizes	included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.; **AND** Composite specifications include **some of the following**: methods for standardizing scales across component scores, scoring rules, weighting rules, handling of missing data, and required sample sizes.	identifying who is included and excluded from the target population, or the event, condition, or outcome being measured; or how to compute the score, etc.; **OR** The composite measure is not well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability **OR** Analysis of missing component scores does not support the specifications for scoring/aggregation and handling of missing component scores. **OR** Analysis not performed
2a2. Reliability testing	Empirical evidence of reliability of **BOTH data elements AND measure score within acceptable norms**. <u>Data element</u>: appropriate method, scope, and reliability statistics for critical data elements within acceptable norms (new testing, or prior evidence for the same data type); **OR** commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission); **OR** *may forego data element reliability testing if data element validity was demonstrated*;	Empirical evidence of reliability **within acceptable norms** for **either critical data elements OR measure score**. <u>Data element</u>: appropriate method, scope, and reliability statistics for critical data elements within acceptable norms (new testing, or prior evidence for the same data type); **OR** commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission); **OR** *may forego data element reliability testing if data element validity was demonstrated*;	Empirical evidence (using appropriate method and scope) of **unreliability** for **either data elements OR measure score,** i.e., statistical results outside of acceptable norms **OR** Inappropriate method or scope of reliability testing

Subcriterion	High	Moderate	Low
	AND **Measure score**: appropriate method, scope, and reliability statistics for score computation or risk adjustment	**OR** **Measure score**: appropriate method, scope, and reliability statistics for score computation or risk adjustment	

2b. Validity

Instructions: To rate "High" for Validity; all subcriteria must have a "High" rating. If there is a combination of "High" and "Moderate" the overall Validity rating is "Moderate". If the measure meets any of the definitions in "Low" column, Validity will be rated "Low" and the measure will fail.

Subcriterion	High	Moderate	Low
2b1. Measure specifications	The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1c) under *Importance to Measure and Report*	*Moderate does not apply. This subcriterion is either rated High or Low*	The measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above;
2b2. Validity testing	Empirical evidence of validity of **BOTH data elements AND measure score within acceptable norms**: **Data element**: appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; **Measure score**: Evidence that supports the intended interpretation of measure scores for the intended purpose—making conclusions about the	Empirical evidence of validity **within acceptable norms for either critical data elements OR measure score** **Data element**: appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; **Measure score**: Evidence that supports the intended interpretation of measure scores for the intended purpose—making conclusions about the	Empirical evidence (using appropriate method and scope) of **invalidity** for **either data elements OR measure score,** i.e., statistical results outside of acceptable norms **OR** Systematic assessment of face validity of measure resulted in **lack of consensus** as to whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and poor quality. **OR** Inappropriate method or scope of

Subcriterion	High	Moderate	Low
	quality of care. Examples of the types of measure score validity testing: Construct validity Discriminative validity/Contrasted groups Predictive validity Convergent validity Reference strategy/Criterion validity	quality of care. Examples of the types of measure score validity testing: Construct validity Discriminative validity/Contrasted groups Predictive validity Convergent validity Reference strategy/Criterion validity **OR** Systematic assessment of face validity of measure, which is the extent to which a measure appears to reflect that which it is supposed to measure "at face value." Face validity for a CMS quality measure may be adequate if accomplished through a systematic and transparent process, by a panel of experts, such as the TEP, where formal rating of the validity is recorded and appropriately aggregated. The TEP should explicitly address whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and poor quality.	validity testing (including inadequate assessment of face validity)
2b2. Validity testing – threats to validity	Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased	*Moderate does not apply. This subcriterion is either rated High or Low*	Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and **determined to bias results** **OR** Threats to validity (lack of risk adjustment/stratification,

Subcriterion	High	Moderate	Low
			multiple data types/methods, systematic missing or "incorrect" data) **are likely and are NOT empirically assessed**
2b3. Exceptions	Exceptions are supported by the clinical evidence, otherwise they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; **AND** Measure specifications for scoring include computing exceptions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); **AND** If patient preference (e.g., informed decision-making) is a basis for exception, there must be evidence that the exception impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exception category computed separately).	*Moderate does not apply. This subcriterion is either rated High or Low*	Exceptions are not supported by evidence, **OR** The effects of the exceptions are not transparent. (i.e., impact is not clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); **OR** If patient preference (e.g., informed decision-making) is a basis for exception, there is no evidence that the exception impacts performance on the measure;
2b4. Risk Adjustment ***For outcome measures and other measures***	An evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are	An evidence-based risk-adjustment strategy is specified consistent with the evaluation criteria, HOWEVER it uses a method that is less than ideal, but acceptable.	A risk-adjustment strategy is not specified AND would be absolutely necessary for the measure to be fair and support valid conclusions about the quality of care.

Subcriterion	High	Moderate	Low
(e.g., resource use) when indicated:	present at start of care, **AND** uses scientifically sound methods; **OR** rationale/data support not using risk adjustment.		
2b5. Meaningful	Data analysis demonstrates that methods for scoring and analysis allow for identification of statistically significant and practically/clinically meaningful differences in performance. **OR** there is evidence of overall less than optimal performance.	Data analysis was conducted but did not demonstrate statistically significant and practically/clinically meaningful differences in performance; HOWEVER the methods for scoring and analysis are appropriate to identify such differences if they exist.	Data analysis was not conducted to identify statistically significant and practically/clinically meaningful differences in performance, **OR** Methods for scoring and analysis are not appropriate to identify such difference.
2b6. Comparable results	***If multiple data sources/methods are allowed*** (specified in the measure), data and analysis demonstrate that they produce comparable results	Multiple data sources/methods are allowed with no formal testing to demonstrate they produce comparable results, HOWEVER there is a credible description of why/how the data elements are equivalent and the results should be comparable	Multiple data sources/methods are allowed and it is unknown if they produce comparable results, **OR** testing demonstrates they do not produce comparable results
2b7. Components fit the conceptual construct;	Component item/measure analysis demonstrates that the included component items/measures fit the conceptual construct; **OR** justification and results for alternative analyses are provided.	Component item/measure analysis demonstrates that the included component items/measures are **somewhat related** to conceptual construct;	Component item/measure analysis **does not demonstrates** that the included component items/measures fit the conceptual construct; **OR** justification and results for alternative analyses is **not provided**.
2b8. Components contribute to the variation	Component item/measure analysis demonstrates that the included components contribute to the variation in the overall	Data analysis was conducted but **did not demonstrate** that the included components contribute to the variation in the overall	Data analysis **was not conducted** to demonstrate that the included components contribute to the variation in the overall composite

Subcriterion	High	Moderate	Low
	composite score; **OR** if not, justification for inclusion is provided	composite score**; HOWEVER** the methods for scoring and analysis are appropriate to identify such differences if they exist	score.
2b9. Aggregation and weighting rules	The scoring/aggregation and weighting rules are consistent with the conceptual construct. If simple, equal weighting is not used, differential weighting **is justified** by empirical analysis or systematic assessment of expert opinion.	The scoring/aggregation and weighting rules are consistent with the conceptual construct. **AND** Simple, equal weighting is not used, and differential weighting **seems to be justified**, and empirical evidence or systematic assessment of expert opinion was not used.	The scoring/aggregation and weighting rules **are not consistent** with the conceptual construct and justification for method is **not provided.**
2b10. Missing component scores	Analysis of missing component scores supports the specifications for scoring/aggregation and handling of missing component scores.	Analysis of missing component scores somewhat supports the specifications for scoring/aggregation and handling of missing component scores.	Analysis of missing component scores does not support the specifications for scoring/aggregation and handling of missing component scores. **OR** Analysis not performed
2c. Disparities			
	High	**Moderate**	**Low**
	Data indicate disparities do not exist; **OR** if disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results	Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities; HOWEVER the rationale/data justify why stratification is neither necessary nor feasible	Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities; **AND** rationale/data are not provided to justify why stratification is neither necessary nor feasible

Guidelines for Summary Rating: Scientific Acceptability of Measure Properties

Instructions: If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 2a, 2b, and 2c to pass this criterion, or NQF will not evaluate it against the remaining criteria. Reliability and validity (2a and 2b) are assessed in combination. In order to determine if the measure passes reliability and validity, use the table below.

Validity Rating	Reliability Rating	Pass	Description
High	Moderate-High	Yes	Evidence of reliability and validity
High	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Moderate	Moderate-High	Yes	Evidence of reliability and validity
Moderate	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Low	Any rating	No	Validity of conclusions about quality is the primary concern. If evidence of validity is rated low, the reliability rating will usually also be low. Low validity and moderate-high reliability represents inconsistent evidence.

Summary Rating: Scientific Acceptability of Measure Properties

Pass: The measure rates **moderate to high on all** aspects of reliability **and** validity **and** disparities

Fail: The measure **rates low** for one or more aspects of reliability **or** validity or disparities

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria for both reliability and validity to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

Feasibility.

Subcriterion	High	Moderate	Low

Subcriterion	High	Moderate	Low
3a Byproduct of care ***(clinical measures only)***	The required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery (e.g., BP reading, diagnosis).	The required data are based on information generated during care delivery; HOWEVER, trained coders or abstractors are required to use the data in computing the measure.	The required data are not generated during care delivery and are difficult to collect or require special surveys or protocols.
3b Electronic data	The required data elements are available in electronic health records or other electronic sources.	If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.	The required data elements are not available in electronic sources **AND** There is no credible, near-term path to electronic collection specified.
3c Inaccuracies, errors, or unintended consequences	Susceptibility to inaccuracies, errors, or unintended consequences related to measurement are judged to be inconsequential or can be minimized through proper actions **OR** can be monitored and detected	There is moderate susceptibility to inaccuracies, errors, or unintended consequences, **AND/OR** They are more difficult to detect through auditing.	Inaccuracies, errors, or unintended consequences have been demonstrated, **AND** They are not easily detected through auditing.
3d Data collection strategy	The measure is in operational use and the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) has been implemented without difficulty.	The measure is not in operational use; **HOWEVER** testing demonstrates the data collection strategy can be implemented with minimal difficulty or additional resources.	The measure is not in operational use, **AND** Testing indicates the data collection strategy was difficult to implement and/or requires substantial additional resources.

Guidelines for Summary Rating: Feasibility

Summary Rating: Feasibility

High rating indicates: **Three or four** subcriteria are rated "**High**".

Moderate rating indicates: "**Moderate" or mixed ratings**, with **no more than one "Low"** rating.

Low rating indicates: **Two or more** subcriteria are rated **"Low".**

Usability and Use

Instructions: If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass both subcriteria 4a, 4b, 4c, 4d, and 4e. A measure must be rated High or Moderate both public reporting and quality improvement usability to pass.

Outcome Measures: An important **outcome** that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement and can be rated "High" or "Moderate".

Subcriterion	High	Moderate	Low
4a. Accountability and Transparency **And** Improvement **And** Benefits of Performance Measure	Measure is currently used in at least one accountability application (public reporting, public health/disease surveillance, payment program, regulatory and accreditation program, professional certification or recognition program, quality improvement with benchmarking, internal quality improvement), **AND** Performance improvement **AND** No unintended consequences were identified during testing **OR** No evidence of unintended consequences to individuals and populations have been reported since implementation.	Measure is not currently used in at least one accountability application and performance improvement; **HOWEVER** there is a credible plan for implementation provided. **AND** A credible rationale that describes how the performance results could be used to further the goal of high quality efficient healthcare for individuals and populations. **AND** Unintended consequences to individuals or populations were identified **AND** Description of how benefits outweigh them were provided or description of actions taken to mitigate them were provided	Credible plan for implementation or rationale for potential use of performance results in quality improvement is **not provided.** **AND** Description of how benefits outweigh them were **not provided** **OR** Description of actions taken to mitigate them were **not provided**

Subcriterion	High	Moderate	Low
4b. Harmonization	The component measure specifications are fully harmonized.	The component measure specifications are **partially harmonized** with related measures HOWEVER the rationale justifies any differences;	The component measure specifications are **not harmonized** with related measures **AND** the rationale does not justify the differences
4c. Distinctive or additive value	Review of existing endorsed measures and measure sets demonstrates that the composite **measure provides a distinctive or additive value** to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of health care, is a more valid or efficient way to measure).	Review of existing endorsed measures and measure sets demonstrates that the composite **measure provides a somewhat distinctive or additive value** to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of health care, is a more valid or efficient way to measure).	Review of existing endorsed measures and measure sets demonstrates that the composite measure **does not provide** a **distinctive or additive value** to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of health care, is a more valid or efficient way to measure).
4d. Can be decomposed	Data detail is maintained such that the composite measure can be decomposed into its components to facilitate transparency and understanding.	*Moderate does not apply. This subcriterion is either rated High or Low*	Data detail is maintained such that the composite measure **cannot be decomposed** into its components to facilitate transparency and understanding.
4e. Achieves stated purpose	Demonstration (through pilot testing or operational data) that the composite measure achieves the stated purpose/objective	Measure has not been tested or operation data is not provided, however, it appears that that the composite measure achieves the stated purpose/objective	It has not been demonstrated (through pilot testing or operational data) that the composite measure achieves the stated purpose/objective and there is no other evidence or rationale provided.

Guidelines for Summary Rating: Usability

Summary Rating: Usability

Pass: The measure rates "M**oderate"** to **"High"** on **all** aspects usability

Fail: The measure rates **"Low"** for **one or more** aspects of usability

5. Harmonization

Instructions: Measures should be assessed for harmonization. Either the measure specifications must be harmonized with related measures so that they are uniform or compatible or the differences must be justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources. Harmonization should not result in inferior measures—measures should be based on the best measure concepts and ways to measure those concepts. There is no presupposition that an endorsed measure is better than a new measure.

Completely	Partially	Not Harmonized
The measure specifications are **completely harmonized** with related measures; the measure can be used at multiple levels or settings/data sources. AND There are no competing measures (already endorsed, in development, or in use)	The measure specifications are **partially harmonized** with related measures, HOWEVER the rationale justifies any differences; the measure can be used at one level or setting/data source	The measure specifications are **not harmonized** with related measures **AND** the rationale does not justify the differences OR There is a competing measure (same focus, same population)

Appendix 4c Tools for Evaluating eMeasures

4c.1 eMeasure

Adapted from National Quality Forum Measure Evaluation Criteria[10]

1. Impact, Opportunity, Evidence—Importance to Measure and Report:

Extent to which the specific measure focus is evidence-based, important to making significant gains in health care quality, and improving health outcomes for a specific high-impact aspect of health care where there is variation in or overall less-than-optimal performance.

1a. High Impact

The measure focus addresses:

- A specific national health goal/priority identified by Department of Health and Human Services (HHS) or the National Priorities Partnership convened by NQF,

OR

- A demonstrated high-impact aspect of health care (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population, leading cause of morbidity/mortality; high resource use (current and/or future), severity of illness, and severity of patient/societal consequences of poor quality).

AND

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement (i.e., data demonstrating considerable variation, or overall less-than-optimal performance, in the quality of care across providers and/or population groups (disparities in care)).

Note: Examples of data on opportunity for improvement include, but are not limited to, prior studies, epidemiologic data, or data from pilot testing or implementation of the proposed measure. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

AND

1c. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- **Health outcome**: a rationale supports the relationship of the health outcome to processes or structures of care.

  Note: Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

- **Intermediate clinical outcome, process, or structure**: A systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measure focus leads to a desired health outcome.

[10] National Quality Forum *Measure Evaluation* Criteria (January 2011). Available at: http://www.qualityforum.org/docs/measure_evaluation_criteria.aspx. Accessed July 31, 2012.

1. Impact, Opportunity, Evidence—Importance to Measure and Report:

Note: Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multi-step process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) **grading definitions** and **methods**, or Grading of Recommendations, Assessment, Development and Evaluation (**GRADE**) **guidelines**.

- **Patient experience with care**: Evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- **Efficiency**: Evidence for the quality component as noted above.

  Note: Measures of efficiency combine the concepts of resource use **and** quality (NQF's *Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures*).

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 1a, 1b and 1c to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

2. Reliability and Validity—Scientific Acceptability of Measure Properties:

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

2a. Reliability

2a1. The measure is well-defined and precisely specified so it can be implemented consistently within and across organizations and allow for comparability, and EHR measure specifications are based on the quality data model (QDM).

The measure specifications (numerator, denominator, exclusions, risk factors) reflect the quality of care problem (1a, 1b) and evidence cited in support of the measure focus (1c) under Importance to Measure and Report.

Crosswalk of the EHR measure specifications (QDM quality data elements, code lists, and measure logic) to the endorsed measure specifications demonstrates that they represent the original measure, which was judged to be a valid indicator of quality.

Note: Measure specifications include the target population (denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window, exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, and scoring/computation.

2a2. Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

Note: eMeasure specifications include data type from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting.

Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to, inter-rater/abstractor or intra-rater/abstractor studies, internal consistency for multi-item scales, and test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

2b. Validity

2. Reliability and Validity—Scientific Acceptability of Measure Properties:
2b1. The measure specifications are consistent with the evidence presented to support the focus of measurement under criterion 1c. The measure is specified to capture the most inclusive target population indicated by the evidence, and exclusions are supported by the evidence.
2b2. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. Data element: Validity demonstrated by analysis of agreement between data elements exported electronically and data elements abstracted from the entire EHR with statistical results within acceptable norms; OR, complete agreement between data elements and computed measure scores obtained by applying the EHR measure specifications to a simulated test EHR data set with known values for the critical data elements. Analysis of comparability of scores produced by the retooled EHR measure specifications with scores produced by the original measure specifications demonstrated similarity within tolerable error limits. Note: Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to, testing hypotheses that the measures scores indicate quality of care (e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method, correlation of measure scores with another valid indicator of quality for the specific topic, or relationship to conceptually-related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; Note: Examples of evidence that an exclusion distorts measure results include, but are not limited to, frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion. **AND** If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). Note: Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
2b4. For outcome measures and other measures when indicated (e.g., resource use): ◆ An evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified, is based on factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care, and has demonstrated adequate discrimination and calibration; **OR** ◆ Rationale/data support no risk adjustment/stratification. Note: Risk factors that influence outcomes should not be specified as exclusions. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race

2. Reliability and Validity—Scientific Acceptability of Measure Properties:
and socioeconomic status rather than to adjust out the differences.
2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance, **OR** There is evidence of overall less-than-optimal performance. Note: With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.
2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.
2c. Disparities If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender); **OR** Rationale/data justifies why stratification is not necessary or not feasible.
(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria for both reliability and validity to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

3. Feasibility: Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement.
3a. For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).
3b. The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.
3c. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified.
3d. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). Note: All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.

4. Usability and Use: Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) are using or cold use performance results for both accountability and performance improvement to achieve the goal of high quality and efficient healthcare for individuals or populations.
4a. Accountability and Transparency Demonstration that performance results of a measure are used or can be used in public reporting, accreditation, licensure, health IT incentives, performance-based payment, or network inclusion/exclusion.
AND **4b. Improvement** Demonstration that performance results facilitate the goal of high quality efficient healthcare or credible rationale that the performance results can be used to further the goal of high quality efficient healthcare. Note: An important outcome that may not have an identified improvement strategy can still be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.
AND **4c.** Benefits of the performance measure in facilitating progress toward achieving high quality efficient healthcare outweigh the evidence of unintended consequences to individuals or populations (if such evidence exists).

5. Harmonization: Extent to which either measure specifications are harmonized with related measures so they are uniform or compatible, or the differences must be justified (e.g. dictated by evidence).
5a. Related measure: The measure specifications for this measure are completely harmonized with a related measure.
5b. Competing measure: This measure is superior to competing measures (e.g. a more valid or efficient way to measure quality); OR has additive value as an endorsed additional measure (provide analyses if possible)

4c.2 eMeasure Evaluation Guidance

(Evaluation criteria adapted from National Quality Forum (NQF) evaluation criteria)

Measure Name:
Measure Set:
Type of Measure:

Instructions: For each subcriterion, check the description that best matches your assessment of the measure. Use the supporting information provided in the Measure Information Form (MIF) and Measure Justification, as well as any additional relevant studies or data. Based on your rating of the subcriteria, use the Measure Evaluation Criteria Summary Rating guidelines included in this tool to make a summary determination for each criterion. Use the information to complete the Measure Evaluation report template.

Importance—Impact, Opportunity, Evidence

Subcriterion	Pass	Fail
1a. High Impact	The measure focus addresses a specific national health goal/priority identified by one or more of the following: ◆ CMS/HHS ◆ Legislative mandate ◆ NQF's National Priorities Partners **OR** The measure focus has high impact on health care as demonstrated by one or more of the following: ◆ Affects large numbers ◆ Substantial impact for a small population ◆ A leading cause of morbidity/mortality ◆ Severity of illness ◆ High Resource Use ◆ Potential cost savings to the Medicare Program (business case[11]) ◆ Patient/societal consequences of poor quality regardless of cost (social case)	The measure does not directly address a national health goal/priority. **AND** The data do not indicate it is a high impact aspect of health care, or is unknown.
1b. Performance Gap	Evidence exists to substantiate a quality problem and opportunity for improvement (i.e., data demonstrate considerable variation) **OR** Data demonstrate overall poor performance across providers or population groups (disparities).	Performance gap is unknown, **OR** There is limited or no room for improvement (no variability across providers or population groups and overall good performance).

Subcriterion	High	Moderate	Low
1c:	5+ studies	2-4 studies	0-1 studies

[11]A business case is described in Section 5, Information Gathering - Appendix 5-C of the Blueprint.

Subcriterion	Pass	Fail	
Quantity of body of evidence: Total number of studies (not articles or papers)			
1c: Quality of body of evidence	Randomized controlled trials (RCTs) of direct evidence, with adequate size to obtain precise estimates of effect, and without serious flaws that introduce bias.	Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect; **OR** RCTs without serious flaws that introduce bias, but with either indirect evidence, or imprecise estimate of effect.	RCTs with flaws that introduce bias **OR** Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations
1c: Consistency of body of evidence	Estimates of benefits and harms to patients are consistent in direction and similar in magnitude across studies in the body of evidence.	Estimates of benefits and harms to patients are consistent in direction, but differ in magnitude across studies in the body of evidence; **OR** If only one study, the estimate of benefits greatly outweighs the estimate of potential harms, **OR** For expert opinion that is systematically assessed, agreement that benefits to patients clearly outweigh potential harms.	Differences in both magnitude and direction of benefits and harms to patients across studies in the body of evidence, or wide confidence intervals prevent estimating net benefit; **OR** For expert opinion evidence that is systematically assessed, lack of agreement that benefits to patients clearly outweigh potential harms.
1c: Potential Exception to Empirical Body of Evidence – (structure and process measures)	*High does not apply. If this exception is applicable, 1c is either rated Moderate or Low*	If there is no empirical evidence, expert opinion is systematically assessed with agreement that the benefits to patients greatly outweigh potential harms.	For expert opinion evidence that is systematically assessed, lack of agreement that benefits to patients clearly outweigh potential harms.
1c: Exception to Empirical Body	Empirical evidence links the health outcome to a known process or structure of care.	A rationale supports the relationship of the health outcome to at least one	No rationale is given that supports the relationship of the health outcome to at least one healthcare

Subcriterion	Pass	Fail	
of Evidence – (health outcome measures)		healthcare structure, process, intervention, or service.	structure, process, intervention, or service.

Guidelines for Summary Rating: Importance

Instructions for evaluating subcriterion 1c Body of Evidence: In order to determine if the measure passes subcriterion 1c body of evidence, use the applicable table below. After evaluating 1c, determine if measure meets Importance by having passed all of the subcriteria (1a, 1b, and 1c).

Structure and Process Measures – rating of body of evidence (1c)

Quantity of Body of Evidence	Quality of Body of Evidence	Consistency of Results of Body of Evidence	Pass Subcriterion 1c
Moderate-High	Moderate-High	Moderate-High	Yes
Low	Moderate-High	Moderate (if only one study, high consistency not possible)	Yes, but only if it is judged that additional research is unlikely to change conclusion that benefits to patients outweigh harms; otherwise, No
Moderate-High	Low	Moderate-High	Yes, but only if it is judged that potential benefits to patients clearly outweigh potential harms; otherwise, No
Low-Moderate-High	Low-Moderate-High	Low	No
Low	Low	Low	No
No empirical evidence ***(potential exception)***	No empirical evidence ***(potential exception)***	No empirical evidence ***(potential exception)***	Yes, but only if it is systematically judged by an expert panel that potential benefits to patients clearly outweigh potential harms; otherwise, No

Health Outcome Measures – exception to rating body of evidence (1c)

Process, Structure or Intervention Linkage to Outcome	Pass Subcriterion 1c
High-Moderate	Yes
Low	No

Summary Rating: Importance

Pass: A**ll** of the subcriteria (1a, 1b, 1c) are rated "**Pass".**
Fail: A**ny** of subcriteria (1a, 1b, 1c) are rated **"Fail".**

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 1a, 1b, and 1c

to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

Scientific Acceptability of Measure Properties —Reliability, Validity, Disparities

2a. Reliability:
Instructions: To rate "High" for reliability; both subcriteria need to have a "High" rating. If there is a combination of "high" and moderate" the overall Reliability rating is "Moderate". If the measure meets any of the definitions in "Low" column, Reliability will be rated "Low" and the measure will fail.

Subcriterion	High	Moderate	Low
2a1. Specifications well defined and precisely specified	All EHR measure specifications are unambiguous and include only data elements from the Quality Data Model (QDM) including quality data elements, code lists, EHR field, measure logic, original source of the data, recorder, and setting; **OR** new data elements are submitted for inclusion in the QDM. Specifications are considered unambiguous if they are likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.	*Moderate does not apply. This subcriterion is either rated High or Low*	One or more EHR measure specifications are ambiguous or do not use data elements from the QDM. Specifications are considered ambiguous if they are not likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.
2a2. Reliability testing	Empirical evidence of reliability of **both data element AND measure score within acceptable norms**: **Data element**: reliability (repeatability) assured with computer programming—**must test data element validity** **AND** **Measure score**: appropriate method, scope, and reliability statistic within acceptable norms	Empirical evidence of reliability **within acceptable norms** for **either data elements OR measure score**. **Data element**: reliability (repeatability) assured with computer programming—**must test data element validity** **AND** **Measure score**: appropriate method, scope, and reliability statistic within acceptable norms	Empirical evidence of **unreliability** for **either data elements OR measure score**—i.e., statistical results outside of acceptable norms

2b. Validity
Instructions: To rate "High" for Validity; all subcriteria must have a "High" rating. If there is a combination of "High" and "Moderate" the overall Validity rating is "Moderate". If the measure meets any of the definitions in "Low" column, Validity will be rated "Low" and the measure will fail.

Subcriterion	High	Moderate	Low
2b1. Measure specifications	The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1c) under *Importance*	*Moderate does not apply. This subcriterion is either rated High or Low*	The measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above;
2b2. Validity testing	Empirical evidence of validity of **both data elements AND measure** score within acceptable norms: **Data element**: validity demonstrated by analysis of agreement between data elements electronically extracted and data elements visually abstracted from the **entire** EHR with statistical results within acceptable norms; **OR** complete agreement between data elements and computed measure scores obtained by applying the EHR measure specifications to a simulated test EHR data set with known values for the critical data elements **Measure score**: appropriate method, scope, and validity testing result within acceptable norms	**Empirical evidence** of validity within acceptable norms for **either data elements OR measure score** **Data element**: validity demonstrated by analysis of agreement between data elements electronically extracted and data elements visually abstracted from the **entire** EHR with statistical results within acceptable norms; **OR** complete agreement between data elements and computed measure scores obtained by applying the EHR measure specifications to a simulated test EHR data set with known values for the critical data elements **Measure score**: appropriate method, scope, and validity testing result within acceptable norms **OR** **Systematic assessment of face validity** of **measure score** as a quality indicator explicitly addressed and found **substantial agreement** that the **scores** obtained **from the measure as specified** will provide an accurate reflection of quality and can be used to distinguish good and poor quality**.**	Empirical evidence (using appropriate method and scope) of **invalidity** for **either data elements OR measure score**, i.e., statistical results outside of acceptable norms **OR** Systematic assessment of face validity of measure resulted in **lack of consensus** as to whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and poor quality. **OR** Inappropriate method or scope of validity testing (including inadequate assessment of face validity)
2b2. Validity testing – threats to validity	Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect"	*Moderate does not apply. This subcriterion is either rated High or Low*	Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect"

Subcriterion	High	Moderate	Low
	data) are empirically assessed and adequately addressed so that results are not biased.		data) are empirically assessed and **determined to bias results** **OR** Threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) **are likely and are NOT empirically assessed**
2b3. Exceptions	Exceptions are supported by the clinical evidence, otherwise they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; **AND** Measure specifications for scoring include computing exceptions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); **AND** If patient preference (e.g., informed decision-making) is a basis for exception, there must be evidence that the exception impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exception category computed separately).	*Moderate does not apply. This subcriterion is either rated High or Low*	Exceptions are not supported by evidence, ◆ **OR** The effects of the exceptions are not transparent. (i.e., impact is not clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); **OR** If patient preference (e.g., informed decision-making) is a basis for exception, there is no evidence that the exception impacts performance on the measure;
2b4. Risk Adjustment ***For outcome measures and other measures (e.g., resource use) when indicated***:	An evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are present at start of care, **AND** uses scientifically sound methods;	An evidence-based risk-adjustment strategy is specified consistent with the evaluation criteria, HOWEVER it uses a method that is less than ideal, but acceptable.	A risk-adjustment strategy is not specified AND would be absolutely necessary for the measure to be fair and support valid conclusions about the quality of care.

Subcriterion	High	Moderate	Low
	OR rationale/data support not using risk adjustment.		
2b5. Meaningful	Data analysis demonstrates that methods for scoring and analysis allow for identification of statistically significant and practically/clinically meaningful differences in performance. **OR** there is evidence of overall less than optimal performance.	Data analysis was conducted but did not demonstrate statistically significant and practically/clinically meaningful differences in performance; HOWEVER the methods for scoring and analysis are appropriate to identify such differences if they exist.	Data analysis was not conducted to identify statistically significant and practically/clinically meaningful differences in performance, **OR** Methods for scoring and analysis are not appropriate to identify such difference.
2b6. Comparable results	***If multiple data sources/methods are allowed*** (specified in the measure), data and analysis demonstrate that they produce comparable results	Multiple data sources/methods are allowed with no formal testing to demonstrate they produce comparable results, HOWEVER there is a credible description of why/how the data elements are equivalent and the results should be comparable	Multiple data sources/methods are allowed and it is unknown if they produce comparable results, **OR** testing demonstrates they do not produce comparable results
For NQF Endorsed measures: *Re-specified* for EHRs – Use this set of ratings to assess reliability and validity of the re-specified measure	The EHR measure specifications use only data elements from the Quality Data Model (QDM) and include quality data elements, code lists, and measure logic; **AND** Crosswalk of the EHR measure specifications (QDM quality data elements, code lists, and measure logic) to the endorsed measure specifications demonstrates that they represent the original measure, which was judged to be a valid indicator of quality; **AND** Analysis of comparability of scores produced by the retooled EHR measure specifications with scores produced by the original measure specifications demonstrated similarity within tolerable error limits.	The EHR measure specifications use only data elements from the QDM as noted above **AND** Crosswalk of the EHR measure specifications as noted above demonstrates that they represent the original measure **AND** ***For measures with time-limited status***, testing of the original measure and evidence ratings of moderate for reliability and validity as described above.	The EHR measure specifications do not use only data elements from the QDM; **OR** Crosswalk of the EHR measure specifications as noted above identifies that they do not represent the original measure **OR** Crosswalk of the EHR measure specifications as noted above was not completed **OR** For measures with time-limited status, empirical evidence of low reliability or validity for original time-limited measure.

2c. Disparities

	High	Moderate	Low

	High	Moderate	Low
	Data indicate disparities do not exist; ◆ **OR** if disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results	Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities; HOWEVER the rationale/data justify why stratification is neither necessary nor feasible	Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities; **AND** rationale/data are not provided to justify why stratification is neither necessary nor feasible

Guidelines for Summary Rating: Scientific Acceptability of Measure Properties

Instructions: If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 2a, 2b, and 2c to pass this criterion, or NQF will not evaluate it against the remaining criteria. Reliability and validity (2a and 2b) are assessed in combination. In order to determine if the measure passes reliability and validity, use the table below.

Validity Rating	Reliability Rating	Pass	Description
High	Moderate-High	Yes	Evidence of reliability and validity
High	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Moderate	Moderate-High	Yes	Evidence of reliability and validity
Moderate	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Low	Any rating	No	Validity of conclusions about quality is the primary concern. If evidence of validity is rated low, the reliability rating will usually also be low. Low validity and moderate-high reliability represents inconsistent evidence.

Summary Rating: Scientific Acceptability of Measure Properties
Pass: The measure rates **moderate to high on all** aspects of reliability **and** validity **and** disparities
Fail: The measure **rates low** for one or more aspects of reliability **or** validity or disparities

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria for both reliability and validity to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

Usability

Instructions: If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass both subcriteria 3a and 3b. A measure must be rated High or Moderate both public reporting and quality improvement usability to pass
Outcome Measures: An important **outcome** that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement and can be rated "High" or "Moderate".

Subcriterion	High	Moderate	Low

Subcriterion	High	Moderate	Low
3a. Public Reporting	Testing demonstrates that information produced by the measure is meaningful, understandable, and useful for **public reporting** (e.g., systematic feedback from users, focus group, cognitive testing).	Formal testing has not been performed, but the measure is in widespread use and you think it is meaningful and understandable for **public reporting** (e.g., focus group, cognitive testing) **OR** *When measure is being rated during its initial development*: A rationale for how the measure performance results will be meaningful, understandable, and useful for public reporting.	The measure is not in use and has not been tested for usability; **OR** Testing demonstrates information produced by the measure is **not meaningful**, understandable, and useful for public reporting **OR** *When measure is being rated during its initial development*: A **rationale** for how the measure performance results will be meaningful, understandable, and useful for public reporting is not provided.
3b. Quality Improvement	Testing demonstrates that information produced by the measure is meaningful, understandable, and useful for **quality improvement** (e.g., systematic feedback from users, analysis of quality improvement initiatives).	Formal testing has not been performed but the measure is in widespread use and accepted to be meaningful and useful for **quality improvement** (e.g., quality improvement initiatives). **OR** *When measure is being rated during its initial development*: A rationale for how the measure performance results will be meaningful, understandable, and useful for quality improvement.	The measure is not in use and has not been tested for usability; **OR** Testing demonstrates information produced by the measure is **not meaningful**, understandable, and useful for public reporting **OR** *When measure is being rated during its initial development*: A **rational**e for how the measure performance results will be meaningful, understandable, and useful for quality improvement is **not provided**.

Guidelines for Summary Rating: Usability

Summary Rating: Usability
Pass: The measure rates "M**oderate"** to **"High"** on **both** aspects usability
Fail: The measure rates **"Low"** for **one or both** aspects of usability

Feasibility.

Subcriterion	High	Moderate	Low
4a. Byproduct of care ***(clinical measures only)***	The required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery (e.g., BP reading, diagnosis).	The required data are based on information generated during care delivery; HOWEVER, the information may not consistently be found in standardized fields and trained coders or abstractors	The required data are not generated during care delivery and are difficult to collect or require special surveys or protocols.

Subcriterion	High	Moderate	Low
		are required to use the data in computing the measure.	
4b. Electronic data	The required data elements are available in electronic health records.	If the required data are not in electronic health records, a credible, near-term path to electronic collection is specified.	The required data elements are not available in electronic sources **AND** There is no credible, near-term path to electronic collection specified.
4c. Inaccuracies, errors, or unintended consequences	Susceptibility to inaccuracies, errors, or unintended consequences related to measurement are judged to be inconsequential or can be minimized through proper actions **OR** can be monitored and detected	There is moderate susceptibility to inaccuracies, errors, or unintended consequences, **AND/OR** They are more difficult to detect through auditing.	Inaccuracies, errors, or unintended consequences have been demonstrated, **AND** They are not easily detected through auditing.
4d. Data collection strategy	The measure is in operational use and the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) has been implemented without difficulty.	The measure is not in operational use; **HOWEVER** testing demonstrates the data collection strategy can be implemented with minimal difficulty or additional resources.	The measure is not in operational use, **AND** Testing indicates the data collection strategy was difficult to implement and/or requires substantial additional resources.

Guidelines for Summary Rating: Feasibility

Summary Rating: Feasibility
High rating indicates: **Three or four** subcriteria are rated "**High**".
Moderate rating indicates: "**Moderate" or mixed ratings**, with **no more than one "Low"** rating.
Low rating indicates: **Two or more** subcriteria are rated **"Low".**

Harmonization

Instructions: Measures should be assessed for harmonization. Either the measure specifications must be harmonized with related measures so that they are uniform or compatible or the differences must be justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources. Harmonization should not result in inferior measures—measures should be based on the best measure concepts and ways to measure those concepts. There is no presupposition that an endorsed measure is better than a new measure.

Completely	Partially	Not Harmonized
The measure specifications are **completely harmonized** with related	The measure specifications are **partially harmonized** with related	The measure specifications are **not harmonized** with related measures

Completely	Partially	Not Harmonized
measures; the measure can be used at multiple levels or settings/data sources	measures, HOWEVER the rationale justifies any differences; the measure can be used at one level or setting/data source	**AND** the rationale does not justify the differences

Appendix 4d Resource Use

Adapted from National Quality Forum Resource Use Measure Evaluation Criteria [12]

1. Impact, Opportunity, Evidence—Importance to Measure and Report:
Resource use measures will be evaluated based on the extent to which the specific measure focus is important to making significant contributions toward understanding health care costs for a specific high-impact aspect of health care where there is variation or a demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, variation in resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality) or overall poor performance.
1a. High Impact The measure focus addresses: ◆ A specific national health goal/priority identified by of the Department of Health and Human Services (HHS) or the National Priorities Partnership convened by NQF; **OR** ◆ A demonstrated high-impact aspect of health care (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).
AND **1b. Performance Gap** Demonstration of resource use or cost problems and opportunity for improvement, i.e., data demonstrating variation in the delivery of care across providers and/or population groups (disparities in care). Note: Examples of data on opportunity for improvement include, but are not limited to, prior studies, epidemiologic data, or data from pilot testing or implementation of the proposed measure. Findings from peer-reviewed literature and empirical data are examples of acceptable information that can be used to justify importance and demonstrating variation. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.
AND **1c.** The purpose/objective of the resource use measure (including its components) and the construct for resource use/costs are clearly described. Note: Measures of efficiency combine the concepts of resource use **and** quality (NQF's *Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures*).
1d. The resource use service categories (i.e., types of resources/costs) that are included in the resource use measure are consistent with and representative of the conceptual construct represented by the measure. Whether or not the resource use measure development begins with a conceptual construct or a set of resource service categories, the service categories

[12] The National Quality Forum. *Resource Use Measure Evaluation Criteria (Version 1.2)*. Available at: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=51459 Accessed: July 31, 2012.

1. Impact, Opportunity, Evidence—Importance to Measure and Report:
included must be conceptually coherent and consistent with the purpose.
(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria 1a, 1b, 1c and 1d to pass this criterion, or NQF will not evaluate it against the remaining criteria.)
2. Reliability and Validity—Scientific Acceptability of Measure Properties:
Extent to which the measure, **as specified**, produces consistent (reliable) and credible (valid) results about the cost or resources used to deliver care.
2a. Reliability **2a1.** The measure is well-defined and precisely specified so it can be implemented consistently within and across organizations and allow for comparability, and EHR measure specifications are based on the quality data model (QDM). Note: Well-defined, complete, and precise specifications for resource use measures include three of the specification modules: measure clinical logic and method, measure construction logic, and adjustments for comparability as relevant to the measure. Data protocol steps are critical to the reliability and validity of the measure; specifications must be detailed enough so that users can execute the necessary steps to implement the measure. Further, additional sub-functions within the data protocol and measure reporting modules may require precise specificity as indicated on the submission form and as appropriate to the submitted measure. To allow for flexibility of measure implementation, clear guidance from the measure developer is required at time of measure submission on those data protocol and measure reporting steps that are not specified with the measure; this guidance will be reviewed for adequacy by the review committees. For those modules and analytic functions that are required in the submission form that the measure developer deems as not relevant or available, justification for and implications of not specifying those steps is required. Specifications should also include the identification of the target population to whom the measure applies, identification of those from the target population who achieved the specific measure focus (i.e., target condition, event), measurement time window, exclusions, risk-adjustment, definitions, data elements, data source and instructions, sampling, and scoring/computation. The resource use measure submission form is the platform through which this information is submitted.
2a2. Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. Note: Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to, inter-rater/abstractor or intra-rater/abstractor studies, internal consistency for multi-item scales, and test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
2b. Validity **2b1.** The measure specifications are consistent with the evidence presented to support the focus of measurement under criterion 1b. The measure is specified to capture the most inclusive target population indicated by the evidence, and exclusions are supported by the evidence.
2b2. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided, adequately distinguishing higher and lower cost and resource use. Note: Validity testing applies to both the data elements and the computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measure's scores indicate resource use, (e.g., measure scores are different for groups known to have differences in resource use

1. Impact, Opportunity, Evidence—Importance to Measure and Report:
assessed by another valid resource use measure or method); correlation of measure scores with another valid indicator of resource use for the specific topic; or relationship to conceptually-related measures. Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish higher from lower resource use or costs. The scoring/aggregation and weighting rules used during measure scoring and construction are consistent with the conceptual construct. If differential weighting is used, it should be justified. Differential weights are determined by empirical analyses or a systematic assessment of expert opinion or value-based priorities. Validity testing for resource use measures can be used to demonstrate validity for each module or the entire measure score. For those modules not included in the demonstration of validity, justification for and implications of not addressing those steps is required.
2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; Note: Examples of evidence that an exclusion distorts measure results include, but are not limited to, frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion. **AND** If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). Note: Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
2b4. For outcome measures and other measures when indicated (e.g., resource use): ◆ An evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; and has demonstrated adequate discrimination and calibration; **OR** ◆ Rationale/data support no risk adjustment/ stratification. Note: Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for cardiovascular disease risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.
2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance, **OR** ◆ There is evidence of overall less-than-optimal performance. Note: With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-

1. Impact, Opportunity, Evidence—Importance to Measure and Report:
than-optimal performance may not demonstrate much variability across providers.
2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.
2c. Disparities If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender); **OR** Rationale/data justifies why stratification is not necessary or not feasible.
(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria for both reliability and validity to pass this criterion, or NQF will not evaluate it against the remaining criteria.)
3. Feasibility:
Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement.
3a. For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).
3b. The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.
3c. Susceptibility to inaccuracies, errors, or unintended consequences related to measurement are judged to be inconsequential or can be minimized through proper actions, OR can be monitored and detected.
3d. The data collection and measurement strategy can be implemented as demonstrated by operational use in external reporting programs, OR testing did not identify barriers to operational use (e.g., barriers related to data availability, timing, frequency, sampling, patient confidentiality, fees for use of proprietary specifications). Note: All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.
4. Usability and Use:
Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) are using or cold use performance results for both accountability and performance improvement to achieve the goal of high quality and efficient healthcare for individuals or populations.
4a. Accountability and Transparency Demonstration that performance results of a measure are used or can be used in public reporting, accreditation, licensure, health IT incentives, performance-based payment, or network inclusion/exclusion.
AND **4b. Improvement** Demonstration that performance results facilitate the goal of high quality efficient healthcare or credible rationale that the performance results can be used to further the goal of high quality efficient healthcare. Note: An important outcome that may not have an identified improvement strategy can still be useful for informing

1. Impact, Opportunity, Evidence—Importance to Measure and Report:
quality improvement by identifying the need for and stimulating new approaches to improvement.
AND **4c.** Benefits of the performance measure in facilitating progress toward achieving high quality efficient healthcare outweigh the evidence of unintended consequences to individuals or populations (if such evidence exists).
4d. Data and result detail are maintained such that the resource use measure, including the clinical and construction logic for a defined unit of measurement can be decomposed to facilitate transparency and understanding.
5. Harmonization: Extent to which either measure specifications are harmonized with related measures so they are uniform or compatible or the differences must be justified (e.g. dictated by evidence).
5a. Related measure: The measure specifications for this measure are completely harmonized with a related measure.
5b. Competing measure: This measure is superior to competing measures (e.g. a more valid or efficient way to measure quality); OR has additive value as an endorsed additional measure (provide analyses if possible)

Appendix 4e Measure Evaluation Report Template

Date as of which information is current:_____

Measure Name:

Measure Set (or setting):

Measure Contractor:

Instructions: *(These instructions are provided for convenience. Delete this section prior to submitting to CMS)*

Middle column: Document the ratings (High, Moderate or Low) for the measure's individual subcriteria. Refer to guidance document

Third column: Provide comments to explain your rating and/or follow-up actions planned for strengthening the subcriteria ratings if rated low or moderate. Include pros/cons, costs/benefits for increasing the rating and the risks if not undertaken.

Summary Rating for each main criteria: Indicate by checking the appropriate box in the Summary Rating sections following the individual subcriteria. Provide a brief statement of conclusions to support the Summary Rating for each of the main criteria.

IMPORTANCE: Impact, Opportunity, Evidence

Subcriteria	Projected NQF Rating	If low (or moderate), plan to increase rating
1a. Impact		
1b. Performance Gap		
1c. Evidence to Support the Measure Focus		

Summary Rating for Importance:

Fail: At least one of the subcriteria above is **not** rated as **high.**

Pass: Measure is important; **all** of the subcriteria are rated **high.**

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

Brief statement of conclusions that support the Summary Rating:

SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES: Reliability and Validity

Subcriteria	Projected NQF Rating	If low (or moderate), plan to increase rating
2a. Reliability		
2a1. Precisely Specified		
2a2. Reliability Testing		
2b. Validity		
2b1. Specifications		
2b2. Validity Testing		
2b3. Exclusions/Exception		
2b4. Risk adjustment		
2b5. Meaningful differences		
2b6. Comparability		
2c. Disparities		

Summary Rating for Scientific Acceptability of Measure Properties:

Pass: The measure rates **moderate to high on all** aspects of reliability **and** validity

Fail: The measure **rates low** for one or more aspects of reliability **or** validity

(If the measure is to be submitted to NQF for endorsement, the measure must be judged to pass all subcriteria for both reliability and validity to pass this criterion, or NQF will not evaluate it against the remaining criteria.)

Rationale for Rating/Comments:

FEASIBILITY

Subcriteria	Projected NQF Rating	If low (or moderate), plan to increase rating
4a. Data a byproduct of care		

Subcriteria	Projected NQF Rating	If low (or moderate), plan to increase rating
4b. Electronic		
4c. Inaccuracies/errors		
4d. Implementation		

Summary Rating for Feasibility:

High rating indicates: The predominant rating for most of the subcriteria is high.

Moderate rating indicates: The predominant rating for most of the subcriteria is moderate.

Low rating indicates: The predominant rating for most of the subcriteria is low.

Rationale for Rating/Comments:

USABILITY and USE

Subcriteria	Projected NQF Rating	If low (or moderate), plan to increase rating
3a.Useful for public reporting		
3a.Useful for quality improvement		

Summary Rating for Usability:

High rating indicates: The predominant rating for most of the subcriteria is high.

Moderate rating indicates: The predominant rating for most of the subcriteria is moderate.

Low rating indicates: The predominant rating for most of the subcriteria is low.

Rationale for Rating/Comments:

Harmonization

Subcriteria	Related or Competing Measures	Plan to harmonize or justification if this measure is superior
5a. Related measure		
5b. Competing measure		

Subcriteria	Related or Competing Measures	Plan to harmonize or justification if this measure is superior
Summary Rating for Harmonization: **High** rating indicates: Measure is completely harmonized with any related measures and there are no competing measures **Moderate** rating indicates: There may be related measures, however, there are justifications for differences; however, there is a some risk that the measure may require further harmonization **Low** rating indicates: There is risk that there may be other measures that are competing or not harmonized with this measure. **Rationale for Rating/Comments:**		

5. Harmonization

5.1 Introduction

Differences in measure specifications limit comparability across settings. Multiple measures with essentially the same focus create confusion in choosing measures to implement and interpreting and comparing the measure results. The National Quality Forum (NQF) requires measure harmonization as part of their endorsement and endorsement maintenance processes, placing it after initial review of the four measure evaluation criteria. Since harmonization should be considered from the very beginning of measure development, the Centers for Medicare & Medicaid (CMS) contractors are expected to consider harmonization as one of the core measure evaluation criteria. CMS has adopted a set of Measure Selection criteria that promotes the use of the same measure or harmonized measures across its programs. Harmonization and alignment work are part of both measure development and measure maintenance. This section highlights points during measure maintenance where the measure contractor should consider harmonization and alignment issues.

Measure harmonization refers to standardization of specifications for:

- Related measures with the same measure focus (e.g., influenza immunization of patients in hospitals or nursing homes).
- Related measures for the same target population (e.g., eye exam and HbA1c for patients with diabetes).
- Definitions applicable to many measures (e.g., age designation for children).

This is done so that specifications are uniform or compatible, unless the differences are justified (e.g., dictated by the evidence).[1]

The dimensions of harmonization can include the numerator, denominator, risk adjustment, exclusions, data source, collection instructions, and other measurement topics. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.[2]

The term “measure alignment” is also used to describe harmonization and related activities. Alignment of measures across programs and agencies and be done at macro and micro levels. Macro-alignment for existing measures consists of agreeing on guiding principles (e.g., outcomes and population-based preferred), the use of standardized measure selection and removal criteria, and similar prioritization of measures and measure concepts across programs and activities.

Micro-alignment of existing measures is similar to harmonization and the terms are sometimes used interchangeably. Micro-alignment may include the standardization of measure specifications (numerator and denominator inclusions and exclusions) across programs, standardization of other

[1]National Quality Forum (NQF), *Guidance for measure harmonization: A consensus report,* Washington, DC: NQF; 2010
[2]Ibid.

aspects of measures and measure sets such as the standardization of value sets from which measures are constructed, standardization of measurement periods used across programs, and standardized, streamlined methods of reporting.

5.2 Procedure

When specifying measures, consider if a similar NQF-endorsed measure or other CMS measure exists for the same condition, process of care, outcome, or care setting. This consideration ensures that the measure under development is harmonized with existing measures.

Measures should be harmonized with existing measures unless there is a compelling reason for not doing so. Harmonization means standardization of specific aspects of similar measures that, when different, do not increase the value or scientific strength of the measure. Harmonization should not result in inferior measures. Measures should be based on the best measure concepts and the best ways to measure those concepts. It should not be assumed that an endorsed measure is better than a new measure.

Harmonizing measure specifications during measure development is more efficient than after a measure has been fully developed and specified. The earlier in the process related or competing measures are identified, the sooner problematic issues may be resolved.

The following includes steps during the measure development process that measure developers can take to achieve measure harmonization.

Step 1: Decide whether to develop a new measure

Conduct an environmental scan for measures already in existence or related measures. Use the information to identify existing measures for the project within a specific topic or condition. If there are no existing or related measures that can be adapted or adopted, then it is appropriate to develop a new measure. (Refer to the Information Gathering section.)

After the Contracting Officer Representative/Government Task Leader (COR/GTL) has approved the recommendations based on the information gathered, develop a set of candidate measures. Work closely with the Measures Manager to ensure that no duplication of measure development occurs. Provide measure development deliverables (candidate lists, etc.) to the Measures Manager, who will help the developer to identify potential harmonization opportunities. Candidate measures may be:

- Existing measures that are adapted.
- Adopted from an existing set of measures.
- Newly developed measures.

Adapted measures

If the measure contractor changes an existing measure to fit the current purpose or use, the measure is considered adapted. This process includes changes to the numerator or denominator specifications,

or revising a measure to meet the needs of a different care setting, data source, or population. Or, it may mean adding additional specifications to fit the current use.

In adapting a measure to a different setting, the measure developer needs to consider accountability, attribution, data source, and reporting tools of the new setting. Measures that are being adapted for use in a different setting or a different unit of analysis may not need to undergo the same level of comprehensive testing or evaluation compared to a newly developed measure. However, when adapting a measure for use in a new setting with a new data source, the newly adapted measure must be evaluated, and possibly re-specified and tested. Before deciding to adapt a measure already in existence, consider the following issues.

- If the existing measure is NQF-endorsed, are the changes to the measure significant enough to require re-submission to NQF for endorsement?
- Will the measure owner be agreeable to the changes in the measure specifications that will meet the needs of the current project?
- If a measure is copyright protected, are there issues relating to the measure's copyright that need to be considered?

Discuss these consideration with the COR/GTL and the measure owner. NQF endorsement status may need to be discussed with NQF. After making any changes to the numerator and denominator statement to fit the particular use, the detailed specifications will be developed.

The first step in evaluating whether or not to adapt a measure is to assess the applicability of the measure focus to the measure topic or setting of interest. Is the measure focus of the existing measure applicable to the quality goal of the new measure topic or setting? Does it meet the importance criterion for the new setting or population?

If the population changes or if the type of data is different, new measure specifications would have to be developed and properly evaluated for soundness and feasibility before a determination regarding use in a different setting can be made. (Refer to the Technical Specifications section for the standardized process.)

Measures that are being adapted for use in a different setting may not need to undergo the same level of development as a new measure. However, aspects of the measure need to be evaluated and possibly re-specified for the new setting in order to show the importance of the measure to each setting the measures may be used for.

Empirical analysis may need to be conducted to evaluate the appropriateness of the measure for the new purpose. The analysis may include, but is not limited to, evaluation of the following:

- Changes in the relative frequency of critical conditions used in the original measure specifications when applied to a new setting/population (e.g., dramatic increase in the occurrence of exclusionary conditions).
- Change in the importance of the original measure in a new setting (e.g., an original measure addressing a highly prevalent condition may not show the same prevalence in a new setting; or,

evidence that large disparities or suboptimal care found based on the original measure may not exist in the new setting/population).

- Changes in the applicability of the original measure (e.g., the original measure composite contains preventive care components that are not appropriate in a new setting such as hospice care).

Adopted measures

Adopted measures must have the same numerator, denominator, and data source as the parent measure. In this case the only information that would need to be provided is particular to the measure's implementation use (such as data submission instructions). If the parent measure is NQF-endorsed and no changes are made to the specifications, the adopted measure is considered endorsed by NQF.

When considering the adoption of an existing measure for use in a CMS program, investigate whether the measure is currently used in another CMS program. CMS generally seeks to align measures across programs unless specific program requirements make this not feasible.

New measures

If the information gathering process and input from the TEP determine that no existing or related measures apply to the topic, then consider a new measure. (Refer to the Information Gathering section.)

Table 5-1 summarizes issues and possible actions to be considered when evaluating measures identified during the environmental scan.[3]

Table 5-1 Harmonization Decisions

	Harmonization Issue	Action
Numerator: Same measure focus Denominator: Same target population	Competing measures	◆ Use existing measure (Adopted), or ◆ Justify development of additional measure. ◆ Different data source will require new specifications that are harmonized.
Numerator: Same measure focus Denominator: Different target population	Related measures	◆ Harmonize on measure focus (Adapted), or ◆ Justify differences, or ◆ Adapt existing measure by expanding the target population.

[3]This table was adapted from The National Quality Forum's *Guidance for Measure Harmonization*, December, 2010, p. 5

	Harmonization Issue	Action
Numerator: Different measure focus Denominator: Same target population	Related measures	◆ Harmonize on target population, or ◆ Justify differences.
Numerator: Different measure focus Denominator: Different target population	New measures	◆ Develop measure.

Step 2: Develop measure specifications

When developing specifications, consider various aspects of the measure for potential harmonization. Harmonization often requires close inspection of specification details of the related measures.

Harmonization may include, but is not limited to:

- Age ranges
- Performance time period
- Allowable values for medical conditions or procedures; code sets, descriptions
- Allowable conditions for inclusion in the denominator; code sets, descriptions
- Exclusion categories, whether exclusions are from the denominator or numerator, whether optional or required
- Calculation algorithm
- Risk adjustment methods

Examples:

- Ambulatory diabetes measures exist, but the new diabetes measure is for a process of care different from existing measures.
- Influenza immunization measures exist for many care settings, but the new measure is for a new care setting.
- Readmission rates exist for several conditions, but the new measure is for a different condition.
- A set of new hospital measures may be able to use data elements already in use for existing hospital measures.

If the measure can be harmonized with any attributes of existing measures, then use the existing definitions for those attributes. Consult with the Measures Manager to review specifications to identify opportunities for further harmonization. If measures should not be harmonized, then document those reasons and include any literature used to support this decision. Some reasons not to harmonize may be that:

- The science behind the new measure does not support using the same variable(s) found in the existing measure.
- CMS's intent for the measure requires the difference.

Examples:

- An existing diabetes measure includes individuals 18–75 years of age. A new process-of-care measure is based on new clinical guidelines that recommend a particular treatment only for individuals older than 65 years of age.
- An existing diabetes measure includes individuals 18–75 years of age. CMS has requested measures for beneficiaries older than 75 years of age.

Consider using data elements that are commonly available in an electronic format within the provider setting. To the extent possible, use existing QDM value sets when developing new eMeasures. The developer should look at the existing library of value sets to determine if there are any that define the clinical concepts described in the measure. If so, use these, rather than create a new value set. This promotes harmonization in addition to decreasing the time needed to research the various code sets to build a new list.

Step 3: Test scientific acceptability of measure properties

For measures that have been adapted, testing may be necessary. Additional testing of the measure in the new setting may also be required for the measure to get NQF endorsement. For further details, please refer to the Measure Testing section.

Step 4: NQF Submission

Although a search for related and competing measures was conducted early during the Information Gathering phase of development prior to submission to NQF, this search should be repeated, as new information may be available. Work closely with the Measures Manager to identify potential related and/or competing measures that may be in development.

There may be instances where the developer creates a new measure when no other measures knowingly exist for the same condition, process of care, outcome, or care setting. However, upon submission to NQF for endorsement, it is noted that similar measures have been submitted. The NQF Steering Committee reviewing the measures can then request that the measure developers create a harmonization plan addressing the possibility and challenges of harmonizing certain aspects of their respective measures. NQF will consider the response and decide whether to recommend the measure for endorsement.

Step 5: Measure selection and implementation

CMS identifies and selects measures it is considering for use. A set of criteria has been adopted by CMS to ensure a consistent approach. Measures selected for use in CMS programs must meet all of the five Core Criteria. Meeting any or all of the four Optional Criteria strengthens the measure's case for being implemented.

The third core criterion addresses the need to have measures that are aligned across CMS and HHS. When considering a measure for topic already measured in another program, it is preferable to use the same measure or a harmonized measure. The CMS Measure Selection Criteria include:

CMS measure selection core criteria

1. Measure addresses an important condition/topic with a performance gap and has a strong scientific evidence base to demonstrate that the measure when implemented can lead to the desired outcomes and/or more appropriate costs (i.e., NQF's Importance criteria).
2. Measure addresses one or more of the six National Quality Strategy Priorities (safer care, effective care coordination, preventing and treating leading causes of mortality and morbidity, person- and family-centered care, supporting better health in communities, making care more affordable).
3. Promotes alignment with specific program attributes and across CMS and HHS programs.
4. Program measure set includes consideration for health care disparities.
5. Measure reporting is feasible.

Optional criteria

1. The measure is responsive to specific program goals or requirements.
2. Enables measurement using a measure type not already measured well (e.g. structure, process, outcome, etc.).
3. Enables measurement across the person-centered episode of care, demonstrated by the assessment of the person's trajectory across providers and settings.
4. Measure set promotes parsimony.

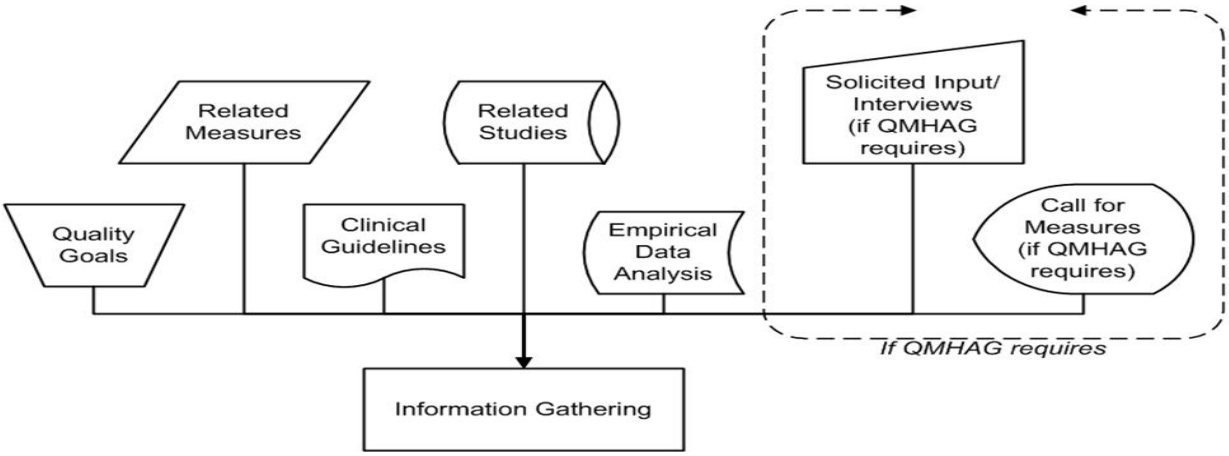
6. Information Gathering

6.1 Introduction

The purpose of this section is to provide guidance for gathering, organizing and documenting information at the initiation of measure development. Information gathering is a broad term that includes an environmental scan (literature review, clinical performance guidelines search, interviews and other related activities) and empirical data analysis. These activities are conducted to obtain information that will guide the prioritization of topics or conditions, gap analysis, business case building, and compilation of existing and related measures. This section describes the various sources of information that can be gathered as well as instructions for documenting and analyzing the collected information.

Information gathering is usually the first step in the development of new measures. Good information gathering will provide a significant knowledge base that includes the quality goals, the strength of scientific evidence (or lack thereof) pertinent to the topics/conditions of interest, as well as information with which to build a business case for the measure. It will also produce evidence of general agreement on the quality issues pertinent to the topics/conditions of interest along with diverse or conflicting views. At a minimum, the four measure evaluation criteria—importance, scientific acceptability of measure properties, feasibility and usability—will serve as a guide for conducting information gathering activities and for identifying priority topics/conditions or measurement areas. The National Quality Forum (NQF) requires measure harmonization as part of their endorsement and endorsement maintenance processes, placing it after initial review of the four measure evaluation criteria. Since harmonization should be considered from the very beginning of measure development, the Centers for Medicare & Medicaid Services (CMS) contractors are expected to consider harmonization as one of the core measure evaluation criteria. Additional criteria that are relevant to the use of the measures may be added by the Contracting Officer Representative/Government Task Leader (COR/GTL). The gathered information will also be used when submitting the measure for NQF endorsement.

Figure 6-1 Information Gathering



6.2 Deliverables

- Information Gathering Report—Summary Report of Environmental Scan and Empirical Analysis
- List of potential candidate measures
- Documentation on Measure Information Form (MIF)
- Documentation on Measure Justification form

6.3 Procedure

Follow the steps below and document your progress using the Information Gathering Report, the Measure Information Form, and the Measure Justification form. The steps are not meant to be sequential but may be undertaken simultaneously.

Step 1: Perform an environmental scan

This includes literature review, clinical practice guidelines review, search for existing and related measures and obtaining stakeholders input to identify important measure topics. Depending upon the nature of the contract and if deemed necessary, the measure contractor may also conduct interviews or post a Call for Measures as part of the environmental scan. While conducting the environmental scan, measure contractors should gather any information necessary to build a business case for the measure and later on document the findings in the Measure Justification form. Examples are available upon request from the Measures Manager.

Literature review

Conduct a literature review to determine the quality issues associated with the topic or setting of interest, and to identify significant areas of controversy if they exist. Document the tools used (e.g., search engines, online publication catalogs) and the criteria (i.e., keywords and Boolean logic) used to conduct the search in the Information Gathering Report Search Methodology section. Whenever possible, include the electronic versions of articles or publications when submitting the report.

Criteria for literature search and required documentation

- Use the measure evaluation criteria. Refer to the Measure Evaluation section to guide the literature search and organize the literature obtained. Specifically, pay attention to:
- Evidence supporting the quality gap associated with the measure topic and the quality of the evidence: This is especially true if (1) clinical practice guidelines are unavailable, (2) there are inconsistent guidelines about the topic, or (3) recent studies have not been incorporated into the guidelines. If recent studies contribute new information that may affect the clinical practice guidelines, the measure contractor must document these studies, even if the measure contractor chooses not to base a measure on the relatively new evidence. Emerging studies or evidence may be an indication that the guideline may change, and if it does, this may affect the stability of the measure.
- Directness of evidence to the specified measure: State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population.

- Quality of the body of evidence: Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors (study design/flaws, directness/indirectness of the evidence to the measure, imprecision/wide confidence intervals due to few patients/events). In general, randomized controlled trials (RCT), studies in which subjects are randomly assigned to various interventions are preferred. However, this type of study is not always available, either because of the strict eligibility criteria or in some case, they may not be appropriate. In these cases non-RCT studies may be relied upon including quasi experimental studies, observational studies (e.g., cohort, case-control, cross-sectional, epidemiological), and qualitative studies.
- Quantity: Five or more RCT studies are preferred. This count refers to actual studies, not papers or journal articles written about the study.
- Consistency of results across studies: Summarize the consistency of direction and magnitude of clinically/practically meaningful benefits over harms to the patients across the studies.
- Grading of strength/quality of the body of evidence: If the body of evidence has been graded, identify the entity that graded the evidence including the balance of representation and any disclosures regarding bias. The measure contractors are not required to grade the evidence, rather, the goal is to assess whether the evidence was graded, and if so, what did the process entail.
- Summary of controversy/contradictory evidence, if applicable.
- Information related to health care disparities in clinical care areas/outcomes across patient demographics: This may include referenced statistics and citations that demonstrate potential disparities (such as race, ethnicity, age, socioeconomic status, income, region, sex, primary language, disability, or other classifications) in clinical care areas/outcomes across patient demographics related to the measure focus. If a disparity has been documented, a discussion of referenced causes and potential interventions should be provided if available.

Sources for literature review

Literature review should include but not be limited to:

- Studies (1) published in peer-reviewed journals, (2) published in journals from respected organizations, (3) written recently (within the last five years), and (4) based on data collected within the last 10 years.
- Unpublished studies or reports such as those described as “gray” literature. Governmental agencies such as the Agency for Healthcare Research and Quality (AHRQ), CMS, and the Centers for Disease Control and Prevention (CDC) produce unpublished studies and reports.
- If available, systematic literature reviews[1] to assess the overall strength of the body of evidence for the measure contract topic. Evaluate each study to grade the body of evidence for the topic.
- Other resources:

[1]A systematic literature review is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research. A systematic review also collects and analyzes data from studies that are included in the review. Two sources of systematic literature reviews are the AHRQ Evidence-Based Clinical Information Reports and The Cochrane Library.

- Institute of Medicine report: *Finding What Works in Health Care Standards for Systematic Reviews*, published March 23, 2011. Available online at http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx
- NQF report: *Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement and Importance to Measure and Report*, published January 2011. Available online at http://www.qualityforum.org/Measuring_Performance/Improving_NQF_Process/Evidence_Task_Force.aspx

Review clinical practice guidelines

Search for the most recent clinical practice guidelines applicable to the measure topic (i.e., written within the past five years). Clinical practice guidelines vary in how they are developed. Guidelines developed by American national physician organizations or federal agencies are preferred. However, guidelines and other evidence documents developed by non-American organizations, as well as non-physician organizations may also be acceptable and should be assessed to determine if they are a sufficient basis for measure development.

Document the criteria used for assessing the quality of the guidelines. When guideline developers use evidence rating schemes, which assign a grade to the quality of the evidence based on the type and design of the research, it is easier for measure contractors to identify the strongest evidence on which to base their measures. If the guidelines were graded, indicate which system was used (United States Preventive Services Task Force (USPSTF) or GRADE).

It is important to note that not all guideline developers use such evidence rating schemes. If no strength of recommendation is noted, document if the guideline recommendations are valid, useful, and applicable.

If multiple guidelines exist for a topic, review the guidelines for consistency of recommendation. If there are inconsistencies among guidelines, evaluate the inconsistencies to determine which guideline will be used as a basis for the measure and document the rationale for selecting one guideline over another.

Sources for clinical practice guidelines review:

- National Guideline Clearinghouse located at http://www.guideline.gov/.
- Institute of Medicine report: Clinical Practice Guidelines We Can Trust, published March 23, 2011. Available online at http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust.aspx.

Search for existing and related measures

Search for measures (existing or related) that will help achieve the quality goals. Keep the search parameters broad to obtain an overall understanding of the measures in existence, including measures that closely meet the contract requirements and other potential sources of information. Look for

measures endorsed and recommended by multi-stakeholder organizations whenever applicable. Include a search for measures developed and/or implemented by the private sector. Determine what types of measures are needed to promote the quality goals for a particular topic/condition or setting. Determine what measurement gaps exist for the topic area, as well as existing measures that may be adopted or adapted for the project. For example, if a contract objective is to development immunization measures for use in the home health setting, it will be necessary to identify and review existing home health measures. In addition, it might also be helpful to analyze immunization measures used in other settings such as nursing homes and hospitals.

The COR/GTL and Measures Management staff can assist in identifying measures in development to ensure that no duplication occurs. Provide measure maintenance deliverables (updated Measure Information Form, updated Measure Justification form, NQF endorsement maintenance documentation, etc.) to the Measures Manager, who will help the developer to identify potential harmonization opportunities. The following table outlines some measure search parameters.

Table 6-1 Measure Search Parameters

Measure Search Parameters
◆ Measures in the same setting, but for a different topic ◆ Measures in a different setting, but for the same topic ◆ Measures that are constructed in a similar manner ◆ Quality indicators ◆ Accreditation standards ◆ NQF preferred practices for the same topic

Search for existing and related measures may involve two steps; searching databases and search for other sources of information, such as performance indicators, accreditation standards, or preferred practices

Search Databases

Use a variety of databases and sources to search for existing and related measures. Below are links to a few readily available sources[2]:

- American Medical Association-Physician Consortium for Performance Improvement—http://www.ama-assn.org/apps/listserv/x-check/qmeasure.cgi?submit=PCPI
- AHRQ National Quality Measures Clearinghouse—http://www.qualitymeasures.ahrq.gov/
- HHS Inventory—http://www.qualitymeasures.ahrq.gov/hhs-measure-inventory/browse.aspx
- National Quality Forum—http://www.qualityforum.org/Measures_List.aspx
- AHRQ—http://www.qualityindicators.ahrq.gv/

[2]The COR/GTL may allow the CMS Measures Manager to share a version of the CMS Measures Inventory with currently used and pipeline measures.

Other sources of information

Search for other sources of information, such as performance indicators, accreditation standards, or preferred practices, that may pertain to the contract topic. Though they may not be as fully developed as a quality measure, quality indicators could be further developed to create a quality measure by providing detailed and precise specifications. Providers seeking accreditation must comply with accreditation standards such as those developed by The Joint Commission or NCQA. Measures aligned with those standards may be easier to implement and be more readily accepted by the providers. NQF endorses preferred practices. These practices are linked to specific desired outcomes, and quality measures may be partially derived from preferred practices.

Stakeholder input to identify measures and important measure topics.

Obtain patient perspective on measure topics

With increasing national focus on a patient–centered approach to improving quality of healthcare; CMS is now requiring measure contractors to seek input from patients to identify healthcare quality issues. Patients can help identify issues that are important to them, and that they find useful to better understand the quality of their care. CMS is making it mandatory to have patients or their caregivers on the TEP. Actual patients or their caregivers should be involved in this process. Participation by consumer and patient advocacy organizations may be desirable but should not be used in lieu of actual patients.

The patients and their caregivers are uniquely qualified to provide input regarding the measure topics or concepts based on their experience. This input should be obtained during Phase 1 TEP discussions. If measure contractors deem appropriate, Phase 1 TEP discussions on importance and usability of measure concepts may be held concurrently with information gathering. Alternatively, these may be held after the information gathering is complete. Refer to the TEP section for further information on Phase 1 discussions.

Interview measure experts, measure developers and other stakeholders (if necessary)

If applicable to the contract, and as directed by the COR/GTL, measure contractor may also contact and interview measure experts, relevant stakeholders, and other measure developers to identify any measures in use or in development relevant to the topic of interest.

Call for measures (if necessary)

While conducting the environmental scan, if insufficient numbers or types of measures have been identified, discuss the situation with COR/GTL to determine if a call for measures is needed. If QMHAG approves, the measure contractor may issue a Call for Measures to the general public. Work with the COR/GTL to develop a list of relevant stakeholder organizations to notify that a Call for Measures is being issued. Measure contractors can notify relevant organizations or individuals about the call for measures before the posting goes live on the Web site. Electronic means can be used to notify the stakeholder community about upcoming calls for measures. Other, more targeted communication can be used to notify relevant stakeholder organizations who can, in turn, notify their members. Relevant stakeholder groups may include but are not limited to quality alliances (AQA, Hospital Quality Alliance

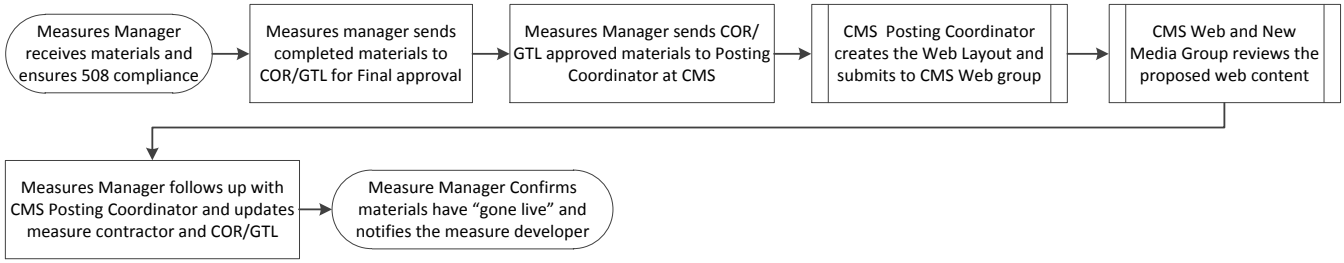
(HQA), PQA, and others), medical societies, scientific organizations related to the measure topic, other CMS measure contractors, etc. In the Call for measures, a measure contractor may request stakeholders to submit candidate measures or measure concepts that meet requirements of the measure contract and if the owner of those measures/measure concepts is willing to expand the measure to CMS. A 14-day call period is recommended.

If an existing measure is found with a measure focus appropriate to the needs of the contract, but the population is not identical, it may be possible for CMS to collaborate with the owner of the original measures to discuss issues related to ownership, maintenance and testing.

Work with the Measures Manager to post the call. Use the "Call for Measures" Web site https://www.cms.gov/MMS/13_CallForMeasures.asp#TopOfPage to document the following:

- Overview of the measure development project.
- Specific project objectives.
- Types of measures needed (outcomes, patient experience of care, care coordination, etc.).
- Instructions on what type of measure information is needed, how long the call for measures will last, etc.
- Contractor's email address where measures and any questions should be sent.

Figure 6-1 The Posting Process



The posting process:

1. After receiving the COR/GTL approved materials, the Measures Manager reviews the materials to confirm they are Section 508 compliant.
2. The Measures Manager sends completed materials to the COR/GTL for final approval and permission to send to CMS Web site Posting Coordinator.
3. The Measures Manager sends the materials to the CMS Web site Posting Coordinator to be loaded into the Web page structure.
4. The CMS Web site Posting Coordinator will create the updated Web page layout and submit it to the CMS Web group for posting.
5. The CMS Web and New Media Group as part of The Office of Communications is responsible for the entire CMS Web site. They will review the proposed Web content to make sure it meets all CMS Web site requirements. Then it is moved to the production environment where the Web page "goes live."

6. The Measures Manager will follow-up with the CMS Web site Posting Coordinator when the approved materials have been moved into the production environment and will update the measure contractor and COR/GTL.
7. The Measures Manager will confirm the materials have "gone live" and will notify the measure developer, the COR/GTL, and the Measures Manager team member working with the contractor.

If a relatively quick turnaround time is required for timely posting, it is best for the contractor to ask the COR/GTL to monitor the process. It is important for measure contractors to understand that the posting process can take up to 5 business days and should be factored into their timeline. Note: materials sent at the end of a business day may not be reviewed until the next business day.

Compile a list of the initial measures received during the call for measures and evaluate these measures using the measure evaluation criteria.

Step 2: Conduct empirical data analysis, as appropriate

If data is available, conduct an empirical data analysis to provide statistical information to support the selection of a topic or condition. The analysis may provide objective data to support the importance of the measure, identify gaps or variations in care; provide incidence/prevalence information, and other data necessary for the development of the business case. This empirical data analysis may also provide quantitative evidence for inclusion or exclusion of a particular set of populations or geographic regions or other considerations for the development of the measure.

Step 3: Evaluate information collected during environmental scan and empirical data analysis

Evaluate existing or related measures for applicability

If there are related measures, evaluate the measures using the measure evaluation criteria to assess if they meet the needs of the measure development contract. If the measure cannot be adapted or adopted to meet the needs of the contract, consider harmonization issues related to similar data elements (e.g., age ranges, time of performance, allowable values for medical conditions or procedures, allowable conditions for inclusion in the denominator, reasons for exclusion from the denominator, risk adjustment methods, etc.). Refer to the Harmonization section for issues related to harmonization considerations.

Regarding adapted measures, if the measure contractor changes an existing measure to fit the current purpose or use, the measure is considered adapted. This may mean changing the numerator or denominator, or changing a measure to meet the needs of a different care setting, data source, or population. Or, it may mean adding additional specifications to fit the current use.

Examples:

- A diabetes measure exists that uses medical claims information; however, the new measure, which is based on the existing measure, uses pharmacy data only.

- A measure for a particular process of care exists for hospitals; however, the new measure is for physician-level use.

The first step in evaluating whether or not to adapt a measure is to assess the applicability of the measure focus to the measure topic or setting of interest. Is the measure focus of the existing measure applicable to the quality goal of the new measure topic or setting? Does it meet the importance criterion for the new setting or population?

If the population changes or if the type of data is different, new measure specifications would have to be developed and properly evaluated for soundness and feasibility before a determination regarding use in a different setting can be made. (Refer to the Technical Specifications section for the standardized process.)

Measures that are being adapted for use in a different setting, the unit of measurement usually do not need to undergo the same level of development as a new measure. However, aspects of the measure need to be evaluated and possibly re-specified for the new setting in order to show the importance of the measure to each setting the measures may be used for. Additional testing of the measure in the new setting may also be required for the measure to get NQF endorsement. For further details, please refer to the Measure Testing section.

Empirical analysis may need to be conducted to evaluate the appropriateness of the measure for the new purpose. The analysis may include, but is not limited to, evaluation of the following:

- Changes in the relative frequency of critical conditions used in the original measure specifications when applied to a new setting/population (i.e., dramatic increase in the occurrence of exclusionary conditions).
- Change in the importance of the original measure in a new setting (i.e., an original measure addressing a highly prevalent condition may not show the same prevalence in a new setting; or, evidence that large disparities or suboptimal care found using the original measure do not exist in the new setting/population).
- Changes in the applicability of the original measure (i.e., the original measure composite contains preventative care components that are not appropriate in a new setting such as hospice care).

If an existing measure is found with a measure focus appropriate to the needs of the contract, but the population is not identical, it may be possible for CMS to collaborate with the owner of the original measures to discuss issues related to ownership, maintenance and testing.

Example:

- A measure for screening adult patients for depression is found. The current contract requires mental health screening measures for adolescents. The owner of the adult depression screening measure may be willing to expand the population in the measure to the adolescent population. If the owner is not willing to expand the population, it may be necessary to develop a new measure specific to the adolescent population which will be harmonized with the existing measure.

- Regarding adopted measures, if the proposed measure has the same numerator, denominator, data source, and care setting as its parent measure, and the only additional information to be provided pertains to the measure's implementation, such as data submission instructions, the measure is considered adopted.

Examples:

- Measures developed and endorsed for physician- or group-level use are specified for submission to a physician group practice demonstration project and are proposed for a new physician incentive program.
- An existing Joint Commission hospital measure not developed by CMS is now added to the CMS measure set.

If a measure is copyright protected, there may be issues relating to its ownership or to proper referencing of the parent measure. In either case, the measure owner will need to be contacted. Upon receiving approval from the original developer to use the existing measures, the detailed specifications will be included for the measure.

Conduct a measurement gap analysis to identify areas for new measure development

Develop a framework to organize the measures gathered. The purpose of this gap analysis is to identify measure types or concepts that may be missing for the measure topic or focus. Refer to Appendix 6a for an example of a framework for assessing needed outcome measures. Through this framework and analysis, the measure contractor may identify existing measures that can be adopted or adapted, or identify the need for new measures that need to be developed.

Determine the appropriate basis for creation of new measures

If there are no existing measures suitable for adoption or adaption and new measures must be developed, the measure contractor will determine the appropriate basis for the new measures based on the material gathered. The appropriate basis will vary by type of measure. It is important to note that the goal is to develop measures most proximal to the outcome desired. Contractors should avoid selecting or constructing measures that can be met primarily through documentation without evaluating the quality of the activity—often satisfied with a checkbox, date, or code—for example a completed assessment, care plan, or delivered instruction.

If applicable to the contract, and as directed by the COR/GTL, the contractor may choose to solicit TEP input to identify the appropriate basis for new measures.

- For process measures, the appropriate basis consists of relevant clinical leverage points. A clinical leverage point is defined as follows:

  > A key state of health, clinical process, or event that has demonstrable effect on the health outcome.Three considerations arise when evaluating leverage points: (1) the area being measured is an important contributing factor to the clinical or contextual process for the goal, (2) the area is one in which measurement and reporting is likely

to stimulate improvement (through either selection or change), and (3) the purpose is to be selective rather than comprehensive." [3]

- For outcome measures, the appropriate basis is the evidence that the specific outcomes of interest are linked to an important CMS goal and may be impacted by clinical interventions. Measures of intermediate outcomes may rely on clinical leverage points as defined above. (Refer to the Special Topics section for more information).
- For structural measures, the appropriate basis is the evidence that the specific structural elements are linked to improved health outcomes or improved care.
- Cost & Resource Use measures should be linked with measures of quality care for the same topic. (Refer to the Special Topics section for more information).
- For all measures, it is important to assess the relationship between the unit of analysis and the decision maker involved. Consider the extent to which processes are under the control of the entity being measured. The measure topic should be attributed to an appropriate provider or setting. This is not an absolute criterion. In some cases there is "shared accountability." For example, for measures of health functioning and care coordination, no one provider controls the performance results.

Examples:

- It is probably not reasonable to hold a physician, who practices in an emergency department (ED), accountable for long-term medication adherence because an ED physician does not provide ongoing management for chronic conditions.
- It may be reasonable to hold an emergency department accountable for timeliness of reperfusion therapy for acute myocardial infarction (AMI) even when the patient is transferred to another hospital. Timeliness relies on the shared accountability of transferring and receiving facilities, as well as the transport team.
- The timely receipt of aspirin when a patient presents with AMI is a shared responsibility of the hospital and emergency department physician and represents an appropriate leverage point for the basis of a measure for both hospitals and emergency department physicians.

Step 4 : Prepare an initial list of measures or measure topics

Develop an initial list of measures based on the results of the previous steps. This list may consist of adopted, adapted or new measures or measure concepts. This list of initial measures should be included in the Information Gathering Report. The measure contractor may document this list of measures or concepts in an appropriate format. One option is to present the measures in a grid or table. This table may include, but is not limited to, the measure name, description, rationale/justification, numerator, denominator, exclusions or exceptions, measure steward, etc. An example of such a grid can be provided upon request. The initial measure list will then be reviewed and measures will be eliminated to create the list of potential measures. Work closely with the Measures Manager to ensure that no duplication of measure development occurs. Provide measure

[3] McGlynn EA. Selecting Measures of Quality and System Performance. *Medical Care*. 2003;41(suppl):39–47.

development deliverables (candidate lists, etc.) to the Measures Manager, who will help the developer to identify potential harmonization opportunities.

Step 5: Apply measure evaluation criteria, and propose a list of potential measures

a) If a large number of measures or concepts were identified, narrow down the list of potential measures by applying the measure evaluation criteria—especially the importance and feasibility criterion to determine which measures should move forward. At a minimum, consider the measure's relevance to the Medicare population; effects on Medicare costs; gaps in care; the availability of well-established, evidence-based clinical guidelines; and/or supporting empirical evidence that can be translated into meaningful quality measures. Other criteria may be included depending on the specific circumstances of the measure set.
b) If applicable to the contract, and as directed by the COR/GTL, the contractor may choose to solicit TEP input to assist in narrowing the list.
c) In the early stages of measure development, while narrowing the initial list of potential measures to candidate measures, the measure developers may use a format such as an MS Excel spreadsheet to present information for multiple measures in one document. Completing the MIF and Measure Justification form should begin as early as possible during the development process. Before presenting measures to the TEP, the developer may choose to use a modified MIF and Measure Justification form to display partial information as it becomes available. At the end of the project, document each potential measure on the MIF and the Measure Justification form.
d) Please note that the MIF and Measure Justification form included in the Measure Development section are aligned with the NQF measure submission form and should be completed in their entirety for new measures or measures that are significantly changed from an original. For measures that are NQF-endorsed and are specified for use in a CMS program, the MIF (excluding the Measure Justification form) may be used.

Step 6: Summarize the evidence in the Measure Justification form

a) Analyze the literature review results and the guidelines found, and organize the evidence to support as many of the measure evaluation criteria as possible. Document the information in the Measure Justification form.
b) Measures that are adopted and NQF-endorsed will not require documentation in the Measure Justification form.
c) The Measure Justification form should be completed for adapted measures. These measures will require evidence of the importance of the topic for new setting or population. The measures may also need to be assessed for reliability and validity, feasibility and usability as well.

Step 7: Submit a report summarizing the information gathered

Prepare a report to the COR/GTL that summarizes the information obtained from the previous steps. This report should include, but not be limited to:

- The methods used to gather information.
- The results of each activity conducted.
- Description of the gap analyses and the results.
- The proposed list of potential measures, along with information gathered for each measure.
- A summary of the information gathered using the Information Gathering Report.

Step 8: Prepare a business case for candidate measures

After the list of potential measures has been presented to the TEP and narrowed to create the list of candidate measures, develop a business case for each candidate measure and document it in the Measure Justification from. Most of the information needed to develop the business case will have been obtained through the initial information gathering. In developing the business case, the measure contractor may need to look for additional information.

Not all topics are associated with financial savings to Medicare; however, a topic may be beneficial for society in general or may be of ethical value. The benefits derived from interventions promoted by the quality measures should be quantified whether in terms of a business case, economic case, or social case. The following types of information should be systematically evaluated to build the business case. Business case examples are available upon request.

- Incidence/prevalence of condition in Medicare population.
- The major benefits of the process or intermediate outcome under consideration for measure (e.g., heart attacks not occurring, hospital length of stay decreased) and the expected magnitude of the benefits.
- Untoward effects of process or intermediate outcome and the likelihood of their occurrence (e.g., bleeding from anticoagulation, death from low blood glucose levels).
- Cost statistics relating to:
  - Cost of implementing the process to be measured.
  - Savings that result from implementing a process.
  - Cost of treating complications that may arise.
- Current process performance, intermediate outcomes, and performance gaps.
- Size of improvement that is reasonable to anticipate.

Development of A Framework for Measurement Gap Analysis

Example[4]

Below is an example of a framework for assessing needed outcomes measures. This framework facilitates selecting outcome measures that include the full scope of services for a cycle of care (moving across the table from left to right), and that assess multiple dimensions of quality for each health system function (moving within columns across the IOM and primary care domains of quality) as required by ACA Section 10303.

Example: Framework for assessing needed outcomes measures

Quality Domain	Population at Risk (Outcomes of Primary Prevention)	Initial Recognition and Management	Management of Acute Event / Surgery	Transitional Care (Post-Acute)	Management of Chronic Disease / Sustaining Health Status
Effective					
Timely					
Efficient					
Safe					
Patient-Centered - Whole person oriented - Sustained partnership - Continuous - Integrated / coordinated					
Equitable					

[4]Centers for Medicare and Medicaid Services. *Selecting Outcomes Measures for Section 10303 of the Patient Protection and Affordable Care Act: A White Paper*. 2011.

Template for Call for Measures Web Page Posting

Project Overview: *<insert title>*

The Centers for Medicare & Medicaid Services (CMS) has contracted with <contractor name> to develop <measure set name or description>. The purpose of the project is to develop measures that can be used to promote quality care to Medicare beneficiaries. As part of its measure development process, CMS may request that interested parties submit candidate or concept measures that may be suitable for this project. If you are submitting fully developed or endorsed measures, attach any additional measure information to the email.

The candidate measures and measure suggestions will be reviewed by CMS, the measure development contractor(s), and Technical Expert Panels convened by the measure contractor. Candidate measures can be access, outcome, process, structure, overuse, underuse, patient experience of care, and care coordination measures. Any measure(s) adapted or adopted for inclusion in the final measure set will reside in the public domain.

Specific objectives include:

- <list contract objectives>

The development process includes:

- Identifying important quality goals related to topic/condition or setting of focus
- Conducting literature reviews and grading evidence
- Defining and developing specifications for each quality measure
- Obtaining evaluation of proposed measures by technical expert panels
- Posting for public comment
- Testing measures for reliability, validity, and feasibility
- Refining measures as needed

Details about the measure development process can be found in the Measures Management System Blueprint https://www.cms.gov/MMS/19_MeasuresManagementSystemBlueprint.asp#TopOfPage.

Instructions:

- Submit the candidate or concept measures on the "Measures Submitted for Consideration" form that is available from the CMS Web site at https://www.cms.gov/MMS/13_CallForMeasures.asp.
- If you are submitting fully developed or endorsed measures, attach any additional measure information to the email.
- Email the completed form and any attachments to: <insert email address>.
- All submissions must be received by *<insert end date for Call for Measures>*.

Measures Submitted for Consideration

Date Submitted

First/Last Name

Suffix (MD, PhD, etc.)/Title

Organization/Mailing Address

Telephone/FAX

E-mail

Candidate measure(s) submitted for consideration (if additional measure information is available, send as an attachment to the e-mail)

Measure Name	Brief Description	Numerator Statement	Denominator Statement	Public Domain Yes/No*	Care Setting	Other Information

*Do you have any commercial, financial or intellectual interest (refer below for definitions) in the candidate measure(s) you are submitting? If yes, please describe the relationship:

- Commercial Interest—A "commercial interest" as defined here, consists of any proprietary entity producing health care goods or services, with the exemption of non-profit or government organizations and non-health care related companies.
- Financial Relationships—Financial relationships are those relationships in which benefits by receiving a salary, royalty, intellectual property rights, consulting fee, honoraria, ownership interest (e.g., stocks, stock options or other ownership interest, excluding diversified mutual funds), or other financial benefit. Financial benefits are usually associated with roles such as employment, management position, independent contractor (including contracted research), consulting, speaking and teaching, membership on advisory committees or review panels, board membership, and other activities from which remuneration is received, or expected. A minimal dollar amount for relationships to be significant has not been set.

Inherent in any amount is the incentive to maintain or increase the value of the relationship. “Relevant” financial relationships in any amount occurring within the past 12 months that create a conflict of interest should be disclosed.

- Intellectual Interest—Intellectual interests may be present when the individual is a principle researcher/investigator in a study that serves as the basis for one or more to the potential performance measure under consideration.

Information Gathering Report—Summary Report of Environmental Scan and Empirical Analysis

Project Name:

Submitter Name:

Date:

Note: If studies and guidelines do not pertain to a measure set in its entirety, but rather are specific to a measure, please indicate that in the summaries.

Search Methodology

Provide a complete explanation of all research tools used, including all online publication directories, keyword combinations, and Boolean logic used to find studies and clinical practice guidelines.

Summary of Literature Review-Studies (Annotated bibliography)

Provide the following information (by individual measure, or, if directed by CMS, provide the information by measure sets):

1. Complete literature citation
2. Level of evidence, rating scheme used
3. Characteristics of the study (population, study size, data sources, study type and methodology)
4. Name which of the four measure evaluation criteria (importance, scientific acceptability, usability, and feasibility) the study addresses. (Sorting the literature review by these criteria will facilitate the development of the measure justification in the later phases of measure development or reevaluation.)
5. Information gathered to build the business case for the measure
   a) Incidence/prevalence of condition in Medicare population.
   b) Identify the major benefits of the process or intermediate outcome under consideration for the measure.
   c) Untoward effects of process or intermediate outcome and likelihood of their occurrence.
   d) Cost statistics relating to cost of implementing the process to be measured; savings that result form implementing process; and cost of treating complications that may arise.
   e) Current performance of process or intermediate outcome and identifying gaps in performance.
   f) Size of improvement that is reasonable to anticipate.
6. Summary of findings
7. Other pertinent information, if applicable

Summary of Clinical Practice Guidelines

Provide the following information (by measure set, or, if needed, provide for individual measures in the set)

1. Guideline name

2. Developer
3. Year published
4. Summary of major recommendations
5. Level of evidence
6. If multiple guidelines exist, note inconsistencies and rationale for using one guideline over another

Measures Gap Analysis

Provide a summary of findings and measurement gaps.

Existing Related Measures

List measures identified.

Measure Name and Description	Numerator and Denominator	Data Source	Care Setting and Unit of Analysis	Developer	Comments[1]

Empirical Data Analysis

For new measures:

1. If available, data source(s) used
2. Time period
3. Methodology
4. Findings

For a measure reevaluation contract, refer to the Comprehensive Reevaluation Form.

1. Obtain current performance data on each measure
2. Analyze measure performance to identify opportunities to improve the measure
3. Provide a summary of empirical data analysis findings

[1]Comments may include the potential for adaptation or adoption to meet the measure contractor's requirements.

Solicit Input/Structured Interviews, if applicable

Include, at a minimum:

1. Summarize overall findings from the input received
2. Name of the person(s) interviewed, type of organization(s) represented, date(s) of interview, the interviewee's area of quality measurement expertise, etc.
3. List of interview questions

7. Technical Expert Panel

7.1 Introduction

A technical expert panel (TEP) is a group of stakeholders and experts who provide direction and thoughtful input to the measure contractor on the development and selection of measures for which the contractor is responsible. With increasing national focus on a patient–centered approach to improving quality of healthcare; CMS is now requiring the measure contractors to include patient(s) or their caregivers on the TEP. Patients and their caregivers are uniquely qualified to provide input regarding the measure topic and concepts based on their own experience and perspective. Since one of the uses of quality measures is to provide information to patients and their caregivers on the quality of care provided, their perspective on what is important and useful to measure and evaluate is vital and cannot be overlooked.

Convening the TEP is a very important step in measure development. If the TEP is to be held early during the contract period, then the contractor should post a call for the panel as soon as the contract is awarded. The Call for TEP is usually done concurrently with the environmental scan, literature review, and other tasks so that the findings are available to present at the TEP meetings. The TEP process is designed to get balanced, multi-stakeholder input. It also is important because TEP input helps the contractor ensure transparency during measure development.

This section focuses on the steps that should be performed when convening a TEP and conducting the TEP meetings. Though the most common reason to convene a TEP is for measure development, there may be other reasons for a TEP as well. Depending on the purpose, the required deliverables from TEP meetings may vary. However, the basic process that a measure contractor should follow to convene a TEP will remain the same.

For most measure development contracts, the measure developer will convene several TEP meetings, either by teleconference or face-to-face. During Phase 1 of the TEP meeting, the members will be asked to review the results of the environmental scan and clarify measure concepts. They will be asked to evaluate the list of potential measures and narrow them down to candidate measures. During Phase 2 discussions and subsequent meetings, the TEP will be asked to review and comment on the measure specifications developed, review the public comments received on the measures, and evaluate results preliminary testing.

Each TEP shall have at minimum one patient or caregiver on its roster to ensure input on the quality issues that are important to patients. Participation by consumer and patient advocacy organizations may be desirable but should not be used in lieu of actual patients. In addition the TEP should include recognized experts in relevant fields, such as clinicians, statisticians, quality improvement experts, methodologists, and pertinent measure developers. TEP members are chosen based on their expertise, personal experience, diversity of perspectives, background and training.

In addition to developing measures that address measurement gaps, the contractor should keep an overall vision for discerning the breadth of quality concerns and related goals for improvement. The TEP should be directed and encouraged to think broadly about principal areas of concern regarding

quality as they relate to the topic or contract at hand. Finally, at the end of the measure development process, the contractor should be able to show how the recommended measures relate to overall Department of Health and Human Services goals including the National Quality Strategy priorities, Centers for Medicare & Medicaid Services (CMS) programs, and recommendations of the Measure Application Partnership.

7.2 Deliverables

- Call for TEP form
- TEP Nomination/Disclosure/Agreement forms
- List of stakeholders
- TEP Charter
- TEP membership list
- TEP meeting and conference call schedule
- Materials for posting the list of TEP members, meeting dates and locations
- Meeting minutes
- Potential measures presented to the TEP
- Measure Evaluation reports
- Updated Measure Justification form and updated Measure Information Form (MIF)
- TEP Summary report, including list of candidate measures

7.3 Procedure

The following steps should be performed when convening the TEP and conducting the TEP meetings. Updates to the Measure Information form and the Measure Justification may be made after the TEP deliberations. For example, information regarding the importance of the measure can be documented in the Measure Justification form.

The TEP process involves three postings to the dedicated CMS Web site, https://www.cms.gov/MMS/15_TechnicalExpertPanels.asp. These three postings include: Call for TEP nominations (Step 3), posting the TEP members, meeting dates and locations, (Step 7) and the TEP summary report (Step 14). The measure contractors will work with the Measures Manager to achieve these postings. The process of making the web site postings will take up to five working days.

Step 1: Complete the Call for TEP form

Use the Call for TEP form to document the following information.

- Overview of the measure development project
- Overall vision for discerning the breadth of quality concerns and related goals for improvement identified for the setting of care
- Specific objectives
- Measure development processes
- Types of expertise needed (such as topic/subject expert, patient perspective, performance measurement, quality improvement, purchaser perspective, health care disparities)

- Expected time commitment and anticipated meeting dates and locations, including any ongoing involvement that is expected to occur throughout the development process
- Instructions for required information (TEP Nomination/Disclosure form, letter of intent)
- Contractor's email address where TEP nominations and any questions are to be sent

The Contracting Officer Representative/Government Task Leader (COR/GTL) will determine if CMS will reimburse for the TEP's travel arrangements based on the current CMS guidelines.

Step 2: Notify relevant stakeholder organizations

Prior to posting the call, share the list of relevant stakeholder organizations for notification about the Call for TEP nominations with the COR/GTL for review and input. Notify the stakeholder organizations regarding the Call for TEP nominations before the posting goes live or simultaneously with the posting on the dedicated Web site so that all stakeholders receive the same message.

Individuals and organizations should be aware that the persons selected to the TEP represent themselves and not their organization. TEP members will use their experience, training, and perspectives to provide input on the proposed measures.

Relevant stakeholder groups may include, but are not limited to:

- Patients and their caregivers or organizations that might help with recruiting appropriate patients / their caregivers.
- Quality alliances (AQA, PQA, and others).
- Medical societies.
- Scientific organizations related to the measure topic.
- Measure developers (AMA-PCPI, NCQA, The Joint Commission, RAND, etc.).
- Other CMS measure contractors.
- Quality Improvement Organizations (QIOs)/ESRD networks.
- Purchaser groups.
- Impacted provider groups/professional organizations.
- Individuals with quality measurement expertise.
- Individuals with health disparities measurement expertise.

Notification methods may include, but are not limited to:

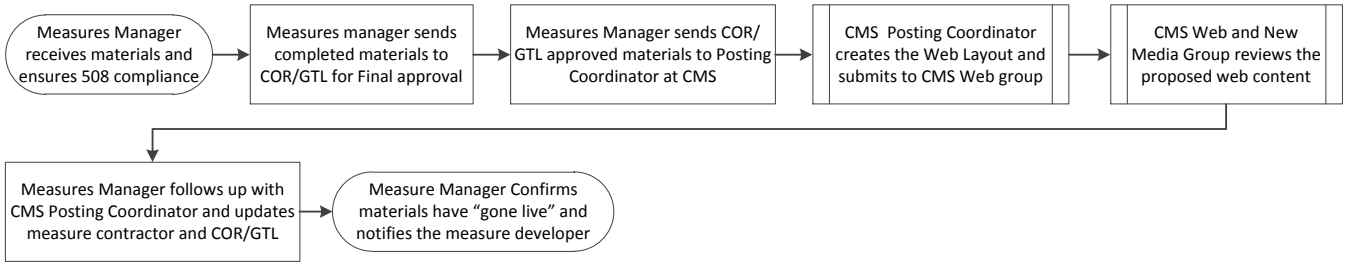
- Sending notice via email to the stakeholders' email lists.
- Emailing to distribution lists and announcing the notification during appropriate CMS workgroup calls.
- Using social media (e.g., Twitter, Facebook, YouTube, LinkedIn, etc.). Contact your COR/GTL for the process.

Step 3: Post the call for nominations following the COR/GTL's approval

Work with the Measures Manager to post the approved Call for TEP form and TEP Nomination/Disclosure/Agreement forms on the dedicated Web site https://www.cms.gov/MMS/15_TechnicalExpertPanels.asp

Refer to the appendices for the template and information required in a Call for TEP and TEP Nomination/Disclosure/Agreement forms.

Figure 7-1 The Posting Process



The posting process:

1. Upon receipt of the posting materials, the Measures Manager reviews to confirm 508 compliance. Information about CMS 508 Compliance is available at http://www.hhs.gov/web/508/index.html.
2. Measures Manager will request final approval from the COR/GTL to send to the CMS Web site Posting Coordinator for posting.
3. Measures Manager sends the materials to the CMS Web site Posting Coordinator for loading.
4. CMS Web site Posting Coordinator creates the updated Web page layout and submits it to the CMS Web group for posting.
5. CMS Web and New Media Group, as part of The Office of Communications, is responsible for the entire CMS Web site. They review the proposed Web content to ensure it meets all CMS Web site requirements. It is then moved to the production environment where the page "goes live."
6. Measures Manager receives confirmation from the CMS Web site Posting Coordinator when the approved materials have been moved into the production environment.
7. Measures Manager confirms that the materials are available on the site and notifies you and the COR/GTL.

If a relatively quick turnaround time is required for timely posting, it is best for you to ask the COR/GTL to monitor the process. The posting process can take up to 5 business days and should be factored into their timeline. Note: materials sent at the end of a business day may not be reviewed until the next business day.

If an insufficient pool of candidates has been received in the call for nominations, the measure contractor should alert the COR/GTL who will need to decide to either:

- Approach relevant organizations or individuals to solicit candidates.
  or
- Choose to extend the call for nominations.

Step 4: Propose the TEP members to the COR/GTL

The average TEP ranges from 8–15 members. This number may be larger or smaller depending on the nature of the contract and level of expertise required. Contracts for multiple measure sets or measures for multiple topics may require multiple TEPs to function simultaneously or within a larger TEP.

Individuals may represent multiple areas of expertise. Consider the following factors when proposing and selecting the TEP members:

- Patient involvement—Actual patients or their caregivers must be integrally involved in this process. Participation by consumer and patient advocacy organizations may be desirable but should not be used in lieu of actual patients. The measure contractor should make every effort to include patient(s) who either have the condition or have experienced the procedure relevant to the measure focus. For example, for hip/knee surgery measures include patient(s) who have undergone or are planning to undergo the surgery. In certain instances it may also be appropriate to include caregivers as TEP members.
- Geography—Include representatives from multiple areas of the country and other characteristics such as rural and urban settings.
- Diversity of experience—Consider individuals with diverse backgrounds and experience in different types of organizations and organizational structures.
- Affiliation—Include members not predominately from any one organization.
- Fair balance—Reasonable effort should be made to have differing points of view represented.
- Availability—select individuals who can commit to attending at least 90 percent of meetings whether they are face-to-face or via telephone. TEP members need to be accessible throughout the performance period of the measure contractor's contract.
- Potential conflict of interest—TEP members are asked to disclose any potential conflict of interest during the nomination process. The contractor should give preference to individuals who will not be inappropriately influenced by any special interest. In the event a person with potential conflict of interest is selected to serve on the TEP, the measure contractor or TEP Chair should inform CMS of the situation. It is for the measure contractor, other TEP members, and the COR/GTL to decide if the individual's interest or relationships may affect the discussions or conclusions. Refer to the TEP Nomination/Disclosure/Agreement form.

Document the proposed TEP member's name and credentials, organizational affiliation, city and state, and area of expertise and experience. Include brief points to clearly indicate why this person was selected. Notify the COR/GTL within one week after the close of the posting about the TEP membership list. Additional information, such as TEP member biographies, may also be sent to the COR/GTL. Confirm each member's participation on the TEP.

Step 5: Select chair and co-chair

Prior to the first TEP meeting, select a TEP chair and co-chair who have either content or measure development expertise. It is important that the TEP chair and co-chair have strong facilitation skills to achieve the following responsibilities:

- Keep the meeting on time
- Conduct the meeting according to the agenda
- Recognize speakers
- Call for votes

Some contractors may choose to bring in an outside facilitator to assist with some of these tasks. However, a TEP chair and co-chair are still required.

The TEP Chair should be available to represent the TEP at the National Quality Forum (NQF) Steering Committee and on follow-up conference calls. However, all TEP members need to be available for possible conference calls with the measure contractor to discuss NQF recommendations.

Step 6: Finalize meeting and conference call schedule

Finalize the meeting and call schedule and inform the COR/GTL.

Step 7: Post the document listing the TEP members, meeting dates and locations

Work with the Measures Manager to post the approved document listing the TEP members. Include the dates and locations of the TEP meetings in the document. Examples are available upon request to the Measures Manager. The information should be available until the TEP Summary report is removed from the Web site.

Step 8: Write TEP Charter

Prepare a document that includes the following information and obtain the COR/GTL's approval:

- TEP name
- Statement that the TEP's role is to provide input to the measure contractor
- Name of contractor convening the TEP
- Scope and objectives of activities
- Overall vision of quality concerns and related goals for improvement
- Description of TEP duties
- Estimated number and frequency of meetings
- TEP composition
- Subgroups (if needed)
- CMS approval date

Examples of TEP charters are available upon request to the Measures Manager.

Step 9: Arrange TEP meetings

Organize and arrange all TEP meetings and conference calls. TEP meetings may occur face-to-face, via telephone conferencing, or a combination of the two. Email communication may also be required. If a face-to-face meeting is required, the measure contractor's staff arranges the meeting, travel and hotel arrangements, meeting rooms, etc.

The measure contractor may decide that additional experts and staff may be needed to support the TEP. The areas of expertise may include, but are not limited to, data management and coding representatives, electronic health records experts, health informatics personnel, and statisticians/health services researchers.

Step 10: Send materials to the TEP

Send the meeting agenda, meeting materials, and supporting documentation to the COR/GTL and TEP members one week prior to the meeting. The measure contractor may also send the TEP charter to members to orient them on their role and responsibilities before the first meeting.

At a minimum, prepare and disseminate the following materials:

- Instructions on the measure evaluation criteria and how the TEP will use them.
- The list of initial or potential measures identified by the measure contractor. Depending on the number of measures that the TEP will review, the contractor, may modify or shorten both the Measure Information form and the Measure Justification form.
  - Measure contractors may modify the MIF to suit their particular contract needs. For example, the contract may not require the developer to develop detailed specifications so a much shorter MIF could be used. Alternatively, contractors who have identified a large number of potential measures may present the measures in a grid or table. This table may include, but is not limited to, the measure name, description, rationale/justification, numerator, denominator, and exclusions.
  - For lay (patients and caregivers) members of the TEP, consideration must be given to present the materials in a manner that they will be able to understand. Patients should not be burdened with detailed technical documents.

Other documents as applicable. All potential TEP members must disclose any current and past activities that may cause a conflict of interest during the nomination process. If at any time while serving on the TEP, a member's status changes and a potential conflict of interest arises, the TEP member is required to notify the measure contractor and TEP Chair. Keep detailed minutes of all TEP meetings whether they are conducted face-to-face or via teleconference. TEP conference calls may be recorded to document the discussion. Announce to the participants if the session is being recorded. At a minimum, the minutes shall include a record of attendance, key points of discussion and input, conclusions reached/decisions made regarding subject matter presented to the TEP, and copies of meeting materials.

Step 11: Conduct the TEP meetings

It is recommended that TEP discussions be held in 2 phases. The goal of Phase 1 is to develop an initial list of measure concepts. Phase 2 is focused on evaluating the measures for further development. Measure contractors can determine the timing of these phases during measure development. For example Phase 1 TEP discussions may be held concurrently during the information gathering phase when collecting information on useful and important measure concepts. Patient participation is mandatory in Phase 1 TEP discussions.

Phase 1 TEP meetings:

This phase should primarily focus on discussions about importance and usability to patients of the potential measures and measure concepts. Patients must participate in these Phase 1 TEP discussions. Measure contractors should particularly be attentive to patients' comments on the collected list of measure ideas as well as to any additional measure ideas and concepts that patient(s) on the TEP might bring during these discussions.

At the opening of the Phase 1 TEP meeting, review and ratify the TEP Charter, explaining the TEP's role and scope of responsibilities. Highlight the role of the patient's voice in measure development. Summarize the findings of the literature review and environmental scan. Discuss any overall quality concerns, such as measurement gaps and alignment across programs and settings as well as overarching goals for improvement. Again, it is important to emphasize that consideration must be given to present the materials in a manner that lay (patients and caregivers) members of the TEP will be able to understand.

By the end of Phase 1 TEP discussions, the measure contractor should be able to identify measures/measure concepts that are deemed important, usable, and valuable by the patient(s) on the TEP which will be discussed further in Phase 2 TEP meetings.

Phase 2 TEP meetings:

Phase 2 TEP meetings should involve in-depth technical discussions about acceptability of the evidence base for the measures, measure specifications, validity, feasibility of the measures etc. Phase 2 technical discussions may be overwhelming for patient TEP members. Patients and caregivers are generally excused at this point. However, they may be included if the contractor or GTL determines it is appropriate to do so. For the Phase 2 TEP meetings, measure contractors should compile a list of measures finalized after Phase 1. Depending on the specifics of the measure contract, measure contractor may seek TEP guidance on one or more measure evaluation criteria based on TEP expertise and as deemed appropriate by the measure contractor. Please refer to the Measure Evaluation section for detailed instructions.

Measure contractors should use the TEP discussions as input to complete the Measure Evaluation report for each measure after the meeting. Alternatively, the measure contractor may conduct a preliminary evaluation of the measures and complete a draft Measure Evaluation report before the TEP meeting. These drafts can be presented to the TEP for discussion.

Work closely with the Measures Manager to ensure that no duplication of measure development occurs. Provide measure development deliverables (candidate lists, etc.) to the Measures Manager, who will help the developer to identify potential harmonization opportunities.

Step 13: Prepare TEP summary report and propose recommended set of candidate measures

Prepare a summary report. At a minimum, the summary will include the following:

- Name of the TEP
- Purpose and objectives of the TEP
- Description of how the measures meet the overall quality concerns and goals for improvement
- Key points of TEP deliberations
- Meeting dates
- TEP composition
- Recommendations on the candidate measures

Step 14: Post the TEP summary report

Work with the Measures Manager to post the approved TEP Summary report— using the posting process detailed in Step 3. It is important for measure contractors to understand that the posting process can take up to 5 business days and should be factored into their timeline. Note: materials sent at the end of a business day may not be reviewed until the next business day.

The report should remain available for at least 21 calendar days or as directed by the COR/GTL.

After the public comment period, the contractor may want to reconvene the TEP to discuss the comments received. Refer to the Public Comment section.

It is important to note that the TEP may be consulted for their advice during any stage of the measure development including when the measure is undergoing the NQF endorsement process.

Template for Call for TEP Web Page Posting

TEP Project Overview: *<insert title>*

The Centers for Medicare & Medicaid Services (CMS) has contracted with *<contractor name>* to develop *<measure set name or description>*. The purpose of the project is to develop measures that can be used to provide quality care to Medicare beneficiaries.

Due date for nominations: *<xx>*

Specific project objectives include:

- <list contract objectives>

The development process includes:

- Identifying important quality goals related to a topic/condition or setting of focus.
- Conducting literature reviews and grading evidence.
- Defining and developing specifications for each quality measure.
- Obtaining evaluation of proposed measures by technical expert panels.
- Posting for public comment.
- Testing measures for reliability, validity, and feasibility.
- Refining measures, as needed.

Details about the measure development process can be found in the Measures Management System Blueprint at https://www.cms.gov/MMS/19_MeasuresManagementSystemBlueprint.asp#TopOfPage.

TEP requirements:

A TEP of approximately *<insert desired TEP size>* individuals will recommend *<insert objective>*. The TEP will be comprised of individuals with the following areas of expertise and perspectives:

- Topic knowledge: <insert specific topic>
- Performance measurement
- Quality improvement
- Patient /Consumer perspective
- Purchaser perspective
- Health care disparities

All potential TEP members must disclose any current and past activities that may pose a potential conflict of interest for performing the tasks required of the TEP. All potential TEP members must also commit to the anticipated time frame needed to perform the functions of the TEP.

TEP expected time commitment:

- <list anticipated meeting dates, locations>
- <if applicable, list expected time frame for measure endorsement activities>

TEP nomination:

Self-nominations are welcome. Third-party nominations must indicate that the individual has been contacted and is willing to serve.

Required information:

- A completed and signed TEP Nomination/Disclosure/Agreement form.

- A letter of interest (not to exceed two pages), highlighting experience/knowledge relevant to the expertise described above and involvement in measure development.
- Curriculum vitae and/or list of relevant experience (e.g., publications), a maximum of 10 pages total.

The TEP Nomination and Disclosure Form can be found in the Download section of https://www.cms.gov/MMS/15_TechnicalExpertPanels.asp#TopOfPage. If you wish to nominate yourself or other individuals for consideration, complete the form and email to: *<insert email address for receipt of the nominations>*.

TEP Nomination/Disclosure/Agreement Form

Project Name: *<Insert Project Name>*

Instructions

Applicants/Nominees must submit the following documents along with this completed and signed form:

- A statement of interest summarizing relevant expertise and knowledge of the applicant (2-page maximum).
- A curriculum vitae (CV) and/or list of relevant experience (e.g., publications) (10-page maximum).
- A disclosure of any current and past activities that may indicate a conflict of interest. As a contractor for CMS, <*measure contractor's name*>, must ensure balance, independence, objectivity and scientific rigor in its measure development activities.
- Send completed and signed form, statement of interest, and CV to <*insert measure contractor name*> with "Nomination" in the subject line at <*insert email address*>. Due by close of business <*insert date*> ET.

All potential TEP members must disclose to the contractor, CMS and other TEP members any significant financial interest or other relationships that may affect their judgment or perceptions. The intent of this disclosure is not to prevent individuals with potential for conflict of interest from serving on the TEP, but to provide the measure contractor, other TEP members, and CMS the information to form their own judgments. It is for the measure contractor, other TEP members, and CMS to decide if the individual's interest or relationships may affect the discussions or conclusions.

Applicant/Nominee Information (Self-nominations Are Acceptable)

First and last name:

Suffix/degrees (RN, MD, PhD, etc.)/Title:

Organization:

Mailing address:

Telephone/fax number(s):

Email address:

Person Recommending the Nominee

Complete this section only if you are nominating a third party for the TEP. You must sign this form and attest that you have notified the nominee of this action and that they are agreeable to serving on the TEP. The measure contractor will request the required information from the nominee.

First and last name:

Suffix (RN, MD, PhD, etc.)/Title:

Organization:

Mailing address:

Telephone/fax number(s):

Email address:

I attest that I have notified the nominee of this action and that he/she is agreeable to serving on the TEP.

Signature: _____ Date: _____

Applicant/Nominee's Disclosure

1. Do you or any family members have a financial interest, arrangement or affiliation with any corporate organizations that may create a potential conflict of interest? *Yes ☐/ No ☐.*

   If yes, please describe (grant/research support, consultant, speaker's bureau, major stock shareholder, other financial or material support). Please include the name of the corporation/organization.

2. Do you or any family members have intellectual interest in a study or other research related to the quality measures under consideration? *Yes ☐/ No ☐.*

   If yes, please describe the type of intellectual interest and the name of the organization/group.

Applicant/Nominee's Agreement

- If at any time during my service as a member of this TEP, my conflict of interest status changes, I will notify the measure contractor and the TEP chair.
- It is anticipated that there will be <*approximate time commitment that is required*>. I am able to commit to attending at least 90 percent of all TEP meetings (face-to-face or by telephone).
- If selected to participate in the TEP and the measures are submitted to a measure endorsement organization (e.g., NQF, AQA) for approval, I will be available to discuss the measures with the organization or its representatives, and work with the measure contractor to make revisions to the measures if necessary.
- If selected to participate in the TEP, I will keep confidential all materials and discussions until such time that CMS authorizes their release.

I have read the above and agree to abide by it.

Signature: _____ Date: _____

Template for Technical Expert Panel Charter

TEP title:

Measure contractor convening the TEP:

Scope and objective of TEP activities:

Description of TEP duties:

Estimated number and frequency of meetings:

Member composition (attach Final List of TEP Members form):

Subgroups (if needed):

Date approved by TEP:

8. Technical Specifications[1]

8.1 Introduction

The purpose of this section is to provide guidance to the measure contractor to ensure that measures developed for Centers for Medicare & Medicaid Services (CMS) have complete technical specifications that are detailed and precise. Final technical specifications provide the comprehensive details that allow the measure to be collected and implemented consistently, reliably, and effectively.

The process of developing measure specifications occurs throughout the measure development process. In the information gathering stage, the measure contractor identifies whether existing measures may be adopted or adapted to fit the desired purpose. If no measures are identified that match the desired purpose for the measure, the contractor must work with its technical expert panel (TEP) to develop new measures. Depending on the information gathering findings, the TEP will consider potential measures that are newly proposed or are derived from existing measures. The measure contractor submits the list of candidate measures, selected with TEP input, to the Contracting Officer Representative/Government Task Leader (COR/GTL) for approval. Upon approval from the COR/GTL, the measure contractor proceeds with the development of detailed technical specifications for the measures.

The Measure Information Form (MIF) is used to document the technical specifications of the measures. These forms may be found in the Measure Development section. At this stage, the technical specifications section is likely to include high-level numerator and denominator statements and initial information on potential exclusions, if applicable, and will continue to be completed throughout the development process as more information is obtained. The fields in the MIF have been aligned with the NQF Measure Submission form, and collecting information on them throughout the development process will make the measure submission process easier. Depending upon the stage of measure development, the MIF can be modified to display only the available information or the developer may use a different format to display multiple measures (i.e., to present to the TEP). The ultimate goal of this process is to collect the information necessary to make the National Quality Forum (NQF) measure submission process easier.

The development of technical specifications is an iterative process. A brief overview of this process is as follows:

- The measure contractor drafts the initial specifications and the TEP will review and advise the measure contractor of the recommended changes. (Refer to the Technical Expert Panel section.)
- If directed by the COR/GTL, public comments may be obtained regarding the draft measures. (Refer to the Public Comment section.) Comments received during the public comment period

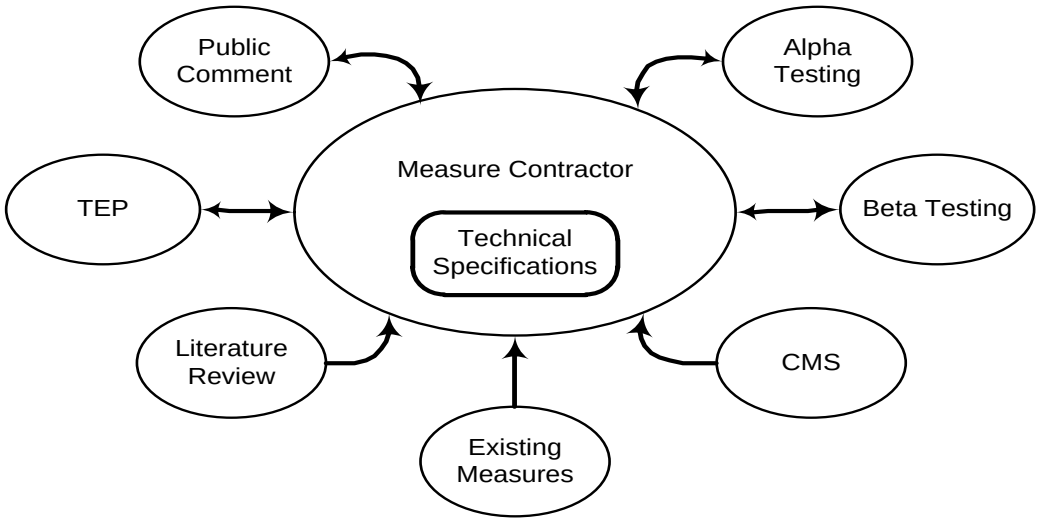
[1]The direction provided in this section is based on guidance from the National Quality Forum and in some instances the verbiage remains unchanged to preserve the intent of the original documents.

will be reviewed and taken into consideration by the measure contractor, CMS, and the TEP and often result in revisions to the measure specifications.

- During the development process, alpha (formative) testing of the measure occurs. For measures based on electronic administrative data (e.g., claims-based), the draft technical specifications may be provided to the programming staff responsible for data retrieval and for developing the programming logic necessary to produce the measure. The programmers will assess the feasibility of the technical specifications as written and may provide feedback. For measures based on chart abstraction, data collection tools are developed and tested.
- When the specifications are more fully developed, beta (field) testing occurs. (Refer to the Measure Testing section.) As a result, technical specifications will continue to evolve and become more detailed and precise.
- After measure testing is complete, additional public comments should be obtained using the formal process outlined in the Public Comment section. All of these factors will result in enhancements to the technical specifications and make the measure more valid and reliable.

Figure 8-1 illustrates the inputs to the process.

Figure 8-1 Factors Influencing the Development of the Measure Technical Specifications



Each measure is required to be precisely specified. Measures must be specified with sufficient details to be distinguishable from other measures and to support consistent implementation across providers.

Most quality measures are expressed as a rate. Usually the basic construct of a measure begins with the numerator, denominator, exclusions, and measure logic. Then the measure concept is more precisely specified with increasing amounts of detail, including the appropriate codes sets and/or detailed and precisely defined data elements.

Figure 8-2 lists key components of the technical specifications used to produce valid and reliable quality measures.

Figure 8-2 Key Components of the Technical Specifications of a Quality Measure

8.2 Deliverables

- Measure Information Form or equivalent document containing the same elements, if approved by the COR/GTL. Refer to the Measure Development Section.
- Measure Justification form (refer to the Measure Development section) to document the measure specifications for new or adapted measures, or measures that are developed and in use by another organization, but are not NQF endorsed.
- The Measure Information Form without the Measure Justification form may be used to document existing measures with NQF endorsement that are being proposed for use by CMS.
  - For contractors developing eMeasures, the Health Quality Measures Format (HQMF) document- which includes a header and a body- should be used in lieu of the MIF. Refer to the eMeasure Specifications section for further details on this format.
- List of potential measures to be developed and timelines

8.3 Procedure

The following steps are to be performed in the development of the measure technical specifications. The different fields within the MIF will be completed and updated as the measure progresses from the information gathering stage to measure testing, and finally NQF endorsement.

Step 1: Develop candidate measure list

Conduct an environmental scan, measure gap analysis, and other information gathering activities to determine if there are existing or related measures before developing new measures. Use the information obtained from the information gathering process to identify existing measures for the project within a specific topic or condition. If there are no existing or related measures that can be

adapted or adopted, then developing a new measure is appropriate. (Refer to the Information Gathering section.)

Provide recommendations based on the results of the environmental scan, measure gap analysis and other information collected during the information gathering process. After the COR/GTL has approved the recommendations, develop a set of candidate measures. Candidate measures may be newly developed measures, adapted existing measures, or measures adopted from an existing set.

Avoid selecting or constructing measures that can be met primarily through documentation without evaluating the quality of the activity (often satisfied with a checkbox, date, or code).

Examples:

- A completed assessment
- A completed care plan
- A delivered instruction (e.g., teaching, counseling)

More important than whether a patient received teaching is whether patients understands how to manage their care, which is best measured from the patients' perspective.[2]

New measures

Begin work on a new measure if it has been determined through the information gathering process and input from the TEP that no existing or related measures are applicable for the topic. (Refer to the Information Gathering section.)

Determine the appropriate basis for measures in consultation with the TEP, keeping in mind the measure evaluation criteria (Refer to the Measure Evaluation section) as a framework. The appropriate basis will vary by type of measure. (Refer to the Information Gathering section.)

- The measure contractor and the TEP draft the measure statement with high-level numerator and denominator statements.
- With input from the TEP, consider the populations to be included in both the numerator and denominator. Also, at this stage, develop a high-level algorithm describing the overall logic that will be used to calculate the measure. Alpha (or formative) testing may be used at this stage to assist with development of the conceptual measure. For measures that are developed using administrative data, data analysis may be conducted to determine strategies for obtaining the desired populations. For measures using medical record information, interviews with clinicians or small-scale tests may assess the feasibility and validity of the measure or portions of the measure. (Refer to the Measure Testing section.)

[2]The National Quality Forum. *CSAC Guidance on Quality Performance Measure Construction*. May 2011. Available at: http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx, Accessed July 31, 2012.

- Consider using existing data elements that are commonly available in an electronic format within the provider setting. When developing measures using electronic health records (EHR), use the Quality Data Model (QDM). (Refer to the eMeasure Specifications section for full discussion of the development of specifications for eMeasures.)
- After determining any areas for potential harmonization (Refer to the Harmonization section), the measure contractor develops the detailed specifications.

Adapted measures

If the measure contractor changes an existing measure to fit the current purpose or use, the measure is considered adapted. This process includes changes to the numerator or denominator specifications, or revising a measure to meet the needs of a different care setting, data source, or population. Or, it may mean adding additional specifications to fit the current use.

In adapting a measure to a different setting, the measure developer needs to consider accountability, attribution, data source, and reporting tools of the new setting. Measures that are being adapted for use in a different setting or a different unit of analysis may not need to undergo the same level of comprehensive testing or evaluation compared to a newly developed measure. However, particularly where the measure is being adapted for use in a new setting with a new data source, this aspect of the adapted measure will need to be evaluated, and possibly re-specified and tested (refer to the Harmonization section for more details). Before the decision is made to adapt a measure in existence, the following issues should be considered:

- If the existing measure is NQF-endorsed, are the changes to the measure significant enough to require resubmission to NQF for endorsement?
- Will the measure owner be agreeable to the changes in the measure specifications that will meet the needs of the current project?
- If a measure is copyright protected, are there issues relating to the measure's copyright that need to be considered?

These considerations must be discussed with the COR/GTL and the measure owner. NQF endorsement status may need to be discussed with NQF. After making any changes to the numerator and denominator statement to fit the particular use, the detailed specifications will be developed.

Adopted measures

For a measure to be considered adopted, it must have the same numerator, denominator, and data source as its parent measure. In this case the only information that would need to be provided is particular to the measure's implementation use (such as data submission instructions). If the parent measure is NQF-endorsed and no changes are made to the specifications, the adopted measure is considered endorsed by NQF**.**

Examples:

- Measures developed and endorsed for physician- or group-level use are specified for submission to a physician group practice demonstration project and are proposed for a new physician incentive program.
- An existing Joint Commission hospital measure not developed by CMS is now added to the CMS measure set.

If a measure is copyright protected, there may be issues relating to its ownership or issues related to proper referencing of the measure. The measure owner may have the right to review and approve any portrayal of the measure before it is disseminated. The measure owner may need to be contacted for details. Measure contractors should beware of the copyright status of any measure proposed for adoption. Before contacting the measure owner, measure contractors should discuss the situation with their COR/GTL.

Composite measures

Select the component measures to be combined in the composite measure. (Refer to the Special Topics section for more details on composite measures.)

Step 2: Develop precise technical specifications and update the MIF

Development of the complete technical specifications is an iterative process. Alpha or formative testing should be conducted, as needed, concurrently with the development of the technical specifications. The timing and types of tests performed may vary depending on variables such as data source, complexity of measures, and whether the measure is new, adapted, or adopted. At a minimum, measures should be specified with the broadest applicability (target population, setting, level of measurement/analysis) as supported by the evidence.[3]

As a starting point, use specifications that exist in the NQF-endorsed measures database or QDM when available. For eMeasure development, the NQF-developed Measure Authoring Tool (MAT) should be used.

Develop measure name and description

Measure name or title:

Briefly convey as much information as possible about the measure focus and target population—abbreviated description.

Format—[target population] who received/had [measure focus]

Examples:

- Patients with diabetes who received an eye exam.

[3] The National Quality Forum. *CSAC Guidance on Quality Performance Measure Construction*. May 2011. Available at: http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx, Accessed July 31, 2012.

- Long-stay residents with a urinary tract infection.
- Adults who received a BMI assessment.

Measure description

Briefly describe the type of score (e.g., percentage, proportion, number) and the target population and focus of measurement.

Format—Patients in the target population who received/had [measure focus] {during [time frame] if different than for target population}

Examples:

- Percentage of residents with a valid target assessment and a valid prior assessment whose need for more assistance with daily activities has increased.
- Median time from emergency department arrival to administration of fibrinolytic therapy in ED patients with ST-segment elevation or left bundle branch block (LBBB) on the electrocardiogram (ECG) performed closest to ED arrival and prior to transfer.
- Adherence to Chronic Medications in individuals over 18 years of age with diabetes.

Define the initial population

The *initial patient population* refers to all patients to be evaluated by a specific performance measure who share a common set of specified characteristics within a specific measurement set to which a given measure belongs.

If the measure is part of a measure set, the broadest group of population for inclusion in the set of measures is the initial population. The cohort from which the denominator population is selected must be specified. Details often include information based upon specific age groups, diagnoses, diagnostic and procedure codes, and enrollment periods. The codes or other data necessary to identify this cohort, as well as any sequencing of steps that are needed to identify cases for inclusion, must also be specified.

Example:

- The population of the AMI measure set is identified using four data elements: ICD-9-CM principal diagnosis code; admission date; birth date; and discharge date. Patients admitted to the hospital for inpatient acute care with an ICD-9-CM principal diagnosis code for AMI as defined in Appendix A, Table 1.1, a patient age (admission date minus birth date) greater than or equal to 18 years and a length of stay (discharge date minus admission date) less than or equal to 120 days are included in the AMI initial patient population and are eligible to be sampled.

Define the denominator

The denominator statement describes the population evaluated by the individual measure. It can be the same as the initial patient population or it is a subset of the initial patient population to further constrain the population for the purpose of the measure. The denominator statement should be

sufficiently described so that the reader understands the eligible population or composition of the denominator. Codes should not be used in lieu of words to express concepts. The denominator statement should be precisely defined and include parameters such as:

- Age ranges
- Diagnosis
- Procedures
- Time window
- Other qualifying events

Format—Patients [age] with [condition] in [setting] during [time frame]

Examples:

- Patients (age 18–75) with diabetes in ambulatory care during a measurement year.
- Female patients 65 years of age and older who responded to the survey indicating they had a urinary incontinence problem in the last six months.
- Patients 18 years of age and older who received at least a 180-day supply of digoxin, including any combination products, in any care setting during the measurement year.
- All patients 18 years of age and older with a diagnosis of chronic obstructive pulmonary disease (COPD) who have a forced expiratory volume in 1 second/forced vital capacity (FEV1/FVC) of less than 70 percent and have symptoms.
- Patients on maintenance hemodialysis during the last hemodialysis treatment of the month, including patients on home hemodialysis.
- All patients 65 years of age and older discharged from any inpatient facility (e.g., a hospital, skilled nursing facility, or rehabilitation facility) and seen within 60 days following discharge in the office by the physician providing ongoing care.

Define the numerator

The numerator statement describes the process, condition, event, or outcome that satisfies the measure focus or intent. Numerators are used in proportion and ratio measures only, and should be precisely defined and include parameters such as:

- The event or events that will satisfy the numerator requirement.
- The performance period or time window in which the numerator event must occur, if it is different from that used for identifying the denominator.

Format—Patients in the target population who received/had [measure focus] {during [time frame] if different than for target population}

Examples:

- Patients in the denominator who received a foot exam including visual inspection, sensory exam with monofilament, or pulse exam.

- Patients with an acute myocardial infarction who had documentation of receiving aspirin within 24 hours before emergency department arrival or during their emergency department stay.
- Nursing home residents in the pneumococcal vaccination sample who had an up-to-date pneumococcal vaccination within the six-month target period as indicated on the selected Minimum Data Set target record (assessment or discharge).

Determine if denominator exclusions or denominator exceptions are allowable

Identify patients who are in the denominator (target population), but who should not receive the process or are not eligible for the outcome for some other reason, particularly where their inclusion may bias results. Exceptions allow for the exercise of clinical judgment, and imply that the treatment was at least considered for each potentially eligible patient. They are most appropriate when contraindications to drugs or procedures being measured are relative, and patients who qualify for one of these exceptions may still receive the intervention after the physician has carefully considered the entire clinical picture.[4] For this reason, most measures apply exceptions only to cases where the numerator is not met.

Example of an exception allowing for clinical judgment:

- COPD is an allowable exception for the use of beta blockers for patients with heart failure; however, physician judgment may determine there is greater benefit for the patient to receive this treatment for heart failure than the risk of a problem occurring due to the patient's coexisting condition of COPD.

Exceptions should be specifically defined where capturing the information in a structured manner fits the clinical workflow. Allowable reasons fall into three general categories**:** medical reasons, patient reasons, and system reasons.

- Medical reasons should be precisely defined and evidence-based. The events excluded should occur often enough to distort the measure results if they are not accounted for. A broadly defined medical reason, such as "any reason documented by physician," may create an uneven comparison if some physicians have reasons that may not be evidence-based.
- Patients' reasons for not receiving the service specified may be an exclusion to allow for patient preferences. Caution needs to be exercised when allowing this type of exclusion.
- System reasons are generally rare. They should be limited to identifiable situations that are known to occur.

Examples:

- Pregnancy: the medication specified in the numerator is shown to cause harm to the fetus (medical reason).

[4]Spertus JA, Bonow RO, Chaun P, et al. ACCF/AHA New Insights Into the Methodology of Performance Measurement: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures. *Circulation.* 2010; 122:2091-2106.

- The patient has a religious conviction that precludes the patient from receiving the specified treatment (patient reason).
- A vaccine shortage prevented administration of the vaccine (system reason).

The exceptions must be captured by explicitly defined data elements that allow analysis of the exceptions to identify patterns of inappropriate exceptions and gaming, and to detect potential health care disparity issues. Analysis of rates without attention to exception information has the potential to mask disparities in health care and differences in provider performance.

Examples:

- A notation in the medical record indicates a reason for not performing the specified care and the reason is not supported by scientific evidence (inappropriate exception).
- Patient refusal may be an exception; however, it has the potential to be overused. For example, a provider does not actively encourage the service, then uses patient refusal as the reason for nonperformance (gaming).
- Patient reason exceptions for mammograms are noted to be high for a particular minority population. This may indicate a need for a more targeted patient education (disparity issues).

The exceptions can sometimes be reported as numerator positives instead of being removed from the denominator. This is sometimes done to preserve denominator size when there is an issue of small numbers. To ensure transparency, the allowable exception (either included as numerator positives or removed from the denominator) must be captured in a way that they could be reported separately, in addition to the overall measure rate.

Denominator exclusions refer to criteria that result in removal from the measure population and denominator before determining if numerator criteria are met. Exclusions are absolute, meaning that the treatment is not applicable and would not be considered. Missing data should not be specified as an exclusion. Missing data may be indicative of a quality problem in itself, so excluding those cases may present an inaccurate representation of quality. Systematic missing data (e.g., if poor performance is selectively not reported) also affects the validity of conclusions that can be made about quality.[5]

Example of an exclusion:

- Patients with bilateral lower extremity amputations from a measure of foot exams.

An allowable exclusion or exception must be supported by:

- Evidence of sufficient frequency of occurrence such that the measure results will be distorted without the exclusions.
- Evidence that the exception is clinically appropriate to the eligible population for the measure.
- Evidence that the exclusion significantly impacts the measure results.

[5]The National Quality Forum. *CSAC Guidance on Quality Performance Measure Construction*. May 2011. Available at: http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx, Accessed July 31, 2012.

Format—patients in the [target population] who [have some additional characteristic, condition, procedure]

There is a significant amount of discussion on the use of exclusions and exceptions, particularly the ability to capture exceptions in electronic health records. There is no agreed upon approach; however additional discussion on the use of exceptions in eMeasures is provided in more detail in the eMeasures Specifications section.

Define the data source(s) to be used in the measure

The source of the data used to calculate a measure has an impact on the reliability, validity, and feasibility of the measure. In addition to the method of data collection that can be used, the specifications should include the sources of data that are acceptable. This may be defined by the contract or be determined by the measure contractor. If more than one data source can be used for the measure, detailed specifications must be developed for each data source. Evidence for comparability of the results calculated from the different data sources must be collected.

Example:

- Breast Cancer Screening. The measure can be calculated from claims, paper medical records, or EHRs. Complete specifications need to be developed and tested for each data source.

Define key terms, data elements and code lists

Terms used in the numerator or denominator statement, or in allowable exclusions/exceptions need to be precisely defined. Some measures are constructed by using precisely defined components or discrete pieces of data often called data elements. If a measure is specified for use in an EHR, these data elements are defined by the Quality Data Model (QDM). QDM elements should be used when available. Should necessary quality data elements not be available in the QDM, the developer should describe and present them to NQF to be considered for addition to the QDM. Technical specifications include the “how” and “where” to collect the required data elements, and measures should be fully specified including all applicable definitions and codes. Precise specifications are essential for implementation.

Example:

- Up-to-date vaccination status—the type of vaccinations to be assessed need to be clearly defined along with the definition of “up-to-date.”

Medical record data from EHRs (for eMeasures, or measures specified for use in an EHR) consist of patient-level information coded in such a way that it can be extracted in a format that can be used in a measure. (Refer to the eMeasure Specifications section for more information.) Information that is captured by a medical records system electronically, but is not coded in such a way that a computer program can extract the information, and thus requires manual abstraction, is not considered electronic data.

Medical record data from paper charts and EHRs *(if not specified for an EHR)* will require instructions for abstraction. The level of detail may require specifying allowable terms, allowable places in the record, and the allowable values.

Examples:

- Allowable terms that can be used from the record—hypertension, high blood pressure, HTN, ↑BP.
- Allowable places within the record—problem list, history and physical, progress notes.
- Allowable values—Systolic BP < 130, Urine dipstick result +1 or greater.

Claims data will require information regarding type of claim, data fields, code types and lists of codes.

Example:

- The AMI mortality measure includes admissions for Medicare FFS beneficiaries aged ≥65 years discharged from non-federal acute care hospitals having a principal discharge diagnosis of AMI and with a complete claims history for the 12 months prior to the date of admission. ICD-9-CM code 410.xx, excluding those with 410.x2 (AMI, subsequent episode of care)

Include in the details of the denominator, numerator and exclusions/exceptions sufficient information so that each person collecting data for the measure will interpret the specifications in the same manner. If multiple data collection methods are allowed, detailed specifications must be produced for each method.

Describe the level of measurement/analysis

The unit of measurement/analysis is the primary entity upon which the measure is applied. The procedure for attributing the measure should be clearly stated and justified. Measures should be specified with the broadest applicability (target population, setting, level of measurement/analysis) as supported by the evidence. However, a measure developed for one level may not be valid for a different level.

Examples:

- A measure created to measure performance by a facility such as a hospital may or may not be valid to measure performance by an individual physician.
- If a claims-based measure is being developed for Medicare use and the literature and guidelines support the measure for all adults, consider not limiting the data source to “Medicare Parts A and B claims.”
- Medication measures developed for use in populations (state or national level), Medicare Advantage plans, prescription drug plans, and individual physician and physician groups.

Describe sampling

If sampling is allowed, describe the sample size or provide guidance in determining the appropriate sample size. Any prescribed sampling methodologies need to be explicitly described.

Determine risk adjustment

Risk adjustment is the statistical process used to identify and adjust for differences in patient characteristics (or risk factors) before examining outcomes of care. The purpose of risk adjustment is to facilitate a more fair and accurate comparison of outcomes of care across health care organizations. (Refer to the Risk Adjustment section.)

Statistical risk models should not include factors associated with disparities of care. Including factors associated with disparities in statistical risk models obscures quality problems related to disparities.[6]

All measure specifications, including the risk adjustment methodology, are to be fully disclosed. The risk adjustment method, data elements, and algorithm are to be fully described in the Measures Information Form and the Risk Adjustment Methodology Report. If calculation requires database-dependent coefficients that change frequently, the existence of such coefficients and the general frequency that they change should be disclosed, but the precise numerical value assigned need not be disclosed because it varies over time. (Refer to the Risk Adjustment section.)

Clearly define any time windows

Time windows must be stated whenever they are used to determine cases for inclusion in the denominator, numerator, or exclusions. Any index event used to determine the time window is to be stated. Developers should avoid the use of ambiguous semantics when referring to time intervals.

Example:

- Medication reconciliation must be performed within **30 days following hospital discharge.** Thirty days is the time window and the hospital discharge date is the index event.
  - This example illustrates ambiguity in interpretation: “30 days” should be clearly identified as calendar days, business days, etc. within the measure guidelines to prevent unintended variances in reportable data.

Describe how the measure results are scored and reported

Most quality measures produce rates; however, there are other scoring methods such as categorical value, continuous variable, count, frequency distribution, non-weighted score/composite/scale, ratio, and weighted score/composite/scales. A description of the type of scoring used is a required part of the measure and should be accompanied by an explanation of the interpretation of the score:

- Better quality = higher score
- Better quality = lower score
- Better quality = score within a defined interval
- Passing score defines better quality

[6]The National Quality Forum. *CSAC Guidance on Quality Performance Measure Construction*. May 2011. Available at: http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx, Accessed July 31, 2012.

Avoid measures where improvement decreases the denominator population (e.g., denominator – patients who received a diagnostic test; numerator – patients who inappropriately received the diagnostic test. With improvement, fewer will receive the diagnostic test).[7]

If multiple rates or stratifications are required for reporting, state it in the specifications. If the allowable exclusions are included in the numerator, specify the measure to report the overall rate as well as the rate of each exclusion. Consideration should be given to stratification by population characteristics. CMS has a continued interest in identifying and mitigating disparities in clinical care areas/outcomes across patient demographics. Therefore, consideration should be given for stratification to detect potential disparities in care/outcomes among populations related to the measure focus. If results are to be stratified by population characteristics, describe the variables used.

Examples:

- A vaccination measure numerator that includes the following: (1) the patient received the vaccine, (2) the patient was offered the vaccine and declined, or (3) the patient has an allergy to vaccine.
  - Overall rate includes all three numerator conditions in the calculation of the rate.
  - Overall rate is reported along with the percentage of the population in each of the three categories.
  - Overall rate is reported with the vaccination rate. The vaccination rate would include only the first condition, that the patient received the vaccine, in the numerator.
- A measure is to be stratified by population type: race, ethnicity, age, socioeconomic status, income, region, sex, primary language, disability, or other classifications.

Develop the calculation algorithm

The calculation algorithm- sometimes referred to as the performance calculation- is an ordered sequence of data element retrieval and aggregation through which numerator and denominator events or continuous variable values are identified by a measure. The developer must describe how to combine and use the data collected to produce measure results. The calculation algorithm can be either a graphical representation or a text description. A calculation algorithm is required for the Measure Information Form and is an item in the NQF Measure Submission form.

The development of the calculation algorithm should be based on the written description of the measure. If the written description of the measure does not contain enough information to develop the algorithm, additional details should be added to the measure. The algorithm is to be checked for consistency with the measure text. The calculation algorithm will serve as the basis for the development of computer programming to produce the measure results.

[7]The National Quality Forum. *CSAC Guidance on Quality Performance Measure Construction*. May 2011. Available at: http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx, Accessed July 31, 2012.

The eMeasure Specifications section provides explicit in instructions for calculation of proportion eMeasures and the process to be used to determine individual and aggregate scores. This ensures that all implementers will come up with the same scores, given the same data and same eMeasures.

Step 3: Document the measures and obtain the COR/GTL's approval

Complete the detailed technical specifications, including any additional documents required to produce the measure as it is intended. The complete specifications, including all attachments, are documented in the Measure Information Form and Measure Justification form (located in the Measure Development section).

Information from measure testing, the public comment period, or other stakeholder input may result in the need to make changes to the technical specifications. The measure contractor will work with the TEP to incorporate these changes before submitting the MIFs and Measure Justification to the COR/GTL for approval.

The MIF and Measure Justification form have been aligned with the NQF Measure Submission Form. Both forms were designed to guide the measure contractor throughout the measure development process in gathering the information in a standardized manner. The forms also provide a crosswalk to the fields in the NQF Measure Submission Form to facilitate online information entry if CMS decides to submit the measure for endorsement. If approved by the COR/GTL, an equivalent document that contains the same information/elements as the MIF may be used.

Step 4: Submit measure to NQF for endorsement

Refer to the National Quality Forum Endorsement section for further discussion of this process.

NQF endorses measures only as a part of a larger project to seek consensus standards (measures) for a given health care condition or practice. If CMS decides that the measures developed under a project are to be submitted to the National Quality Forum (NQF) for endorsement, and NQF is conducting a project for which the measure is applicable, the measure contractor will assist CMS in the submission process. Measures are submitted to NQF using the Web-based electronic submission form. Upon the direction of the COR/GTL, the contractor will initiate the submission and complete the submission form.

NQF makes periodic updates to the measure submission process. Consult the NQF Web site below for the current process.

http://www.qualityforum.org/Measuring_Performance/Consensus_Development_Process.aspx

If the measure is submitted to NQF for endorsement, the measure contractor is required to provide technical support throughout the review process. This may include presenting the measure to the Steering Committee that is evaluating the measure, answering questions from NQF staff or the Steering Committee about the specifications, testing or evidence.

During the course of the review, NQF may recommend revisions to the measure. If so, all subsequent revisions must be reported to and approved by the COR/GTL. Update the MIF with any approved changes.

The measure contractor for any measure developed under contract with CMS should list CMS as the steward, unless special arrangements have been made in advance. Developers should consult with the COR/GTL if there are questions on this. Barring special arrangements, the following format should be used—Centers for Medicare & Medicaid Services (CMS).

8.4 Special Considerations

Data source

Administrative data

Electronic data often includes transactional data that has been created for the purpose of billing. This information can come from claims that have been submitted and adjudicated or from the provider's billing system.

Benefit programs categorize Medicare claims as follows:

- Part A is hospital insurance provided by Medicare. Part A covers inpatient care in skilled nursing facilities, critical access hospitals, and hospitals. Hospice and home health care are also covered by Part A.
- Part B covers outpatient care, physician services, physical or occupational therapists, and additional home health care.
- Part D is a stand-alone prescription drug coverage insurance administered by companies offering prescription drug plans.

Claims from each of these Medicare benefits have specific types of information and are unique sources of data containing data elements that can be used in the development of a quality measure. Claims data can be used if CMS or its contractors will calculate the measure results.

Similar data elements may exist in the provider's billing system that can be used to produce claims. This information may be appropriate if the provider is to calculate the measure or identify cases for the denominator.

Other types of administrative data include patient demographics obtained from eligibility or enrollment information, physician office practice management systems, and census information. Payroll data and other databases containing information about providers can also be a source for some types of measures.

Electronic clinical data

Electronic clinical data consists of patient-level information that can be extracted in a format that can be used in a measure. Information that is captured by a medical records system electronically, but is

not coded in such a way that a computer program can extract the information, and thus requires manual abstraction, is not considered electronic data. Electronic health records (EHRs) are one form of electronic clinical data, and these systems often include laboratory, imaging, and pharmacy data that can be queried and extracted for the purpose of the measure. (Refer to the eMeasure Specifications section for more information.)

Medical records (paper-based or electronic)

Medical records are a traditional source of clinical data for measures, and the data may be documented on paper or electronically. This data source, however, requires labor intensive manual extraction. Information manually abstracted from an EHR, which may include clinical laboratory data, imaging services data, personal health records, and pharmacy data, may be used in a quality measure and should be considered the same or similar to a paper patient record.

Registry

The term *registry* can apply to a variety of electronic sources of clinical information that can be used as a data source for quality measures.

In general, a registry is a collection of clinical data for the purposes of assessing clinical performance quality of care. The system records all clinically relevant information about each patient, as well as the population of patients as a whole.

Registries may be components of an EHR of an individual clinician practice, or may be part of a larger regional or national system that may operate across multiple clinicians and institutions. An example of a registry that is part of an EHR of an individual physician or practice is a diabetes registry. This type of registry identifies all of the patients in the practice who have diabetes and tracks the clinical information on this set of patients for this condition.

Examples of national registries include the Action Registry-GWTG (from the American College of Cardiology Foundation and American Heart Association), the Society of Thoracic Surgeons Database, and Paul Coverdell National Acute Stroke Registry. These registries collect data at the facility level. Information can be reported to the facility for local quality improvement or aggregated and reported at a regional or national level.

Registries have been used by public health departments for many years to record cases of diseases of public health importance. This type of registry can provide epidemiological information that can be used to calculate incidence rates and risks, maintain surveillance, and monitor trends in incidence and mortality. Immunization registries are used to collect and maintain complete and current vaccination records to promote disease prevention and control.

Patient assessments

CMS uses data items/elements health assessment instruments or question sets. These assessment instruments or question sets have been developed to provide the requisite data properties to develop and calculate quality measures. Examples of this type of data include the Long Term Care (LTC) Facility

Resident Assessment Instrument (RAI), the Outcome and Assessment Information Set (OASIS), Minimum Data Set (MDS), etc.

Surveys

Surveys are another source of data. Survey data may be collected directly from a patient, such as the Consumer Assessment of Healthcare Providers and Systems (**CAHPS**) surveys that assess the satisfaction and experiences of patients in various settings of care. Other surveys are collected from providers, such as NCQA's Physician Practice Connections. A number of advantages for using survey data exist. Several sources of survey data are readily available, surveys ask about concepts such as satisfaction that are not available elsewhere, and surveys provide a unique window into the patient's feelings.

Code sets (including ICD-10 discussion)

Most CMS measures rely at least in part on the use of various code sets for classifying health care provided in the United States. Any codes that are required for the measure will need to be listed along with their source. Any instructions pertaining to their use need to be explicitly stated. Specifications may require certain codes to be accompanied by certain other codes, or to occur in certain positions, or on claims from specific provider types. Some code sets may require copyright statements to accompany their use. CPT® is a registered trademark of the American Medical Association (AMA). The Current Procedural Terminology (CPT) code sets are owned and maintained by the AMA and require current copyright statements to accompany their use. Logical Observation Identifiers Names and Codes (LOINC), copyrighted by the Regenstrief Institute and Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), are other proprietary code sets that should be properly presented.

Below are some commonly used code sets along with some consideration for their use.

International Classification of Diseases (ICD)

International Classification of Diseases (ICD) is used for identifying data on claims records, data collection for use in performance measurement, and reimbursement for Medicare/Medicaid medical claims. ICD is an epidemiological classification used to identify diagnoses (diseases, injuries, and impairments). The U.S. version also includes procedures (surgical, diagnostic, and therapeutic).

Although the ICD-9-CM Coordination and Maintenance Committee is a federal committee, suggestions for modifications come from both the public and private sectors. Interested parties are asked to submit recommendations for modification prior to a scheduled meeting. Proposals for a new code should include a description of the code being requested, and rationale for why the new code is needed. Supporting references and literature may also be submitted. Proposals should be consistent with the structure and conventions of the classification.

These meetings are open to the public; comments are encouraged both at the meetings and in writing. Recommendations and comments are carefully reviewed and evaluated before any final decisions are made. No decisions are made at the meetings. The ICD-9-CM Coordination and Maintenance Committee's role is advisory. All final decisions are made by the director of National Center for Health

Statistics (NCHS) and the administrator of CMS. Final decisions are made after the December meeting and become effective October 1 of the following year.[8]

On January 16, 2009, the U.S. Department of Health and Human Services (HHS) released a final rule mandating that everyone covered by the Health Insurance Portability and Accountability Act (HIPAA) must implement ICD-10 for medical coding by October 1, 2013. However, on April 17, 2012 HHS published a proposed rule that would delay the compliance date for ICD-10 from October 1, 2013 to October 1, 2014.[9]

The current system, International Classification of Diseases, 9th Edition, Clinical Modification (ICD-9-CM), does not provide the necessary detail for patients' medical conditions or the procedures and services performed on hospitalized patients.

The new classification system provides significant improvements through greater detailed information and the ability to expand in order to capture additional advancements in clinical medicine.

ICD-10-CM/PCS consists of two parts:

- ICD-10-CM—The diagnosis classification system developed by the Centers for Disease Control and Prevention for use in all U.S. health care treatment settings. Diagnosis coding under this system uses 3–7 alpha and numeric digits and full code titles, but the format is very much the same as ICD-9-CM.
- ICD-10-PCS—The procedure classification system developed by the Centers for Medicare & Medicaid Services (CMS) for use only in the U.S. for inpatient hospital settings. The new procedure coding system uses 7 alpha or numeric digits while the ICD-9-CM coding system uses 3 or 4 numeric digits.[10]

As a result of the revised timeline for ICD-10 implementation, NQF also published a revised timeline regarding the requirements for measures using ICD codes:

- October 2011—Measure developers/stewards will be required to submit ICD-9-CM and ICD-10-CM/ PCS codes for review for all endorsement-maintenance projects.
- October 2014—Measure developers/stewards will be required to submit ICD-10-CM/ PCS for HIPAA transactions.
- January 2015—ICD-9-CM codes will no longer be accepted for measure specifications after December 31, 2014.[11]

When a developer submits ICD-10 codes, then the following requirements should also be met:

---

[8] ICD-9-CM Coordination and Maintenance Committee. Available at: http://www.cdc.gov/nchs/icd/icd9cm_maintenance.htm. Accessed July 31, 2012.

[9] Centers for Medicare & Medicaid Services. *ICD-10, Statute and Regulations.* Available at: http://www.cms.gov/ICD10/02d_CMS_ICD-10_Industry_Email_Updates.asp#TopOfPage. Accessed July 31, 2012.

[10] Centers for Medicare & Medicaid Services. *ICD-10 CM/PCS – An Introduction*. Available at: http://www.cms.gov/ICD10/Downloads/ICD-10Overview.pdf. Accessed July 31, 2012.

[11] The National Quality Forum. *Measure Developer Webinar, June 18, 2012.* Available at: http://www.qualityforum.org/Calendar/2012/06/Measure_Developer_Webinar.aspx. Accessed July 31, 2012.

- Provide a statement of intent for the selection of ICD-10 codes, chosen from the following:
  - Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.
  - Goal was to take advantage of the more specific code set to form a new version of the measure, but fully consistent with the original intent.
  - The intent of the measure has changed.

- Provide a spreadsheet, including:
  - A full listing of ICD-9 and ICD-10 codes, with code definitions.
  - The conversion table (if there is one).

- Provide a description of the process used to identify ICD-10 codes, including:
  - Names and credentials of any experts who assisted in the process.
  - Name of the tool used to identify/map to ICD-10 codes.
  - Summary of stakeholder comments received.[12]

Below is the schedule of updates to ICD-9 and ICD-10 Code Sets during the transition period.

- October 1, 2011—Last annual update to ICD-9 and ICD-10. Code set partial freeze began.
- October 1, 2012—Limited updates to ICD-9 and ICD-10 for new technologies (ICD-9 ends after this update).
- October 1, 2013—Claims for services provided on or after this date must use ICD-10 codes for medical diagnosis and inpatient procedures.
- October 1, 2014—Regular updates to ICD-10 code sets begins. Code set partial freeze ends.

Current Procedural Terminology, Fourth Edition (CPT4®)

CPT is a registered trademark of the American Medical Association (AMA) for the Current Procedural Terminology, Fourth Edition (CPT4). The CPT Category I or CPT codes is a listing of descriptive terms and identifying codes for reporting medical services and procedures performed by physicians. The purpose of the terminology is to provide a uniform language that will accurately describe medical, surgical, and diagnostic services, and will thereby provide an effective means for reliable nationwide communication among physicians, patients, and third parties.[13] This code set is updated annually.

Each CPT record corresponds to a single observation or diagnosis. The CPT codes are not intended to transmit all possible information about an observation, or diagnosis. They are only intended to identify the observation or diagnosis. The CPT code for a name is unique and permanent.

---

[12]The National Quality Forum. *ICD-10-CM/PCS Coding Maintenance Operational Guidance: A Consensus Report.* Available at: http://www.qualityforum.org/Publications/2010/10/ICD-10-CM/PCS_Coding_Maintenance_Operational_Guidance.aspx. Accessed July 31, 2012.

[13]Centers for Medicare & Medicaid Services. *Data Submission Specifications Utilizing HL7 QRDA Implementation Guide Based on HL7 CDA Release 2.0 Version: 2.0* Last Modified: July 01, 2010 Available at: https://www.cms.gov/PQRI/20_AlternativeReportingMechanisms.asp#TopOfPage Accessed August 7, 2012.

Current Procedural Terminology, Fourth Edition (CPT®) is copyrighted by the AMA, All Rights Reserved. CPT is a registered trademark of the AMA.

CPT Category II or CPT II codes, developed through the CPT Editorial Panel for use in performance measurement, serve to encode the clinical actions described in a measure's numerator. CPT II codes consist of five alphanumeric characters in a string ending with the letter "F." CPT II codes are updated annually and are not modified or updated during the year.

When publishing measures that use CPT codes, users must include a set of notices and disclosures required by the AMA. Contact the COR/GTL to obtain the current full set of notices and disclaimers that includes:[14]

- Copyright notice
- Trademark notice
- Government rights statement
- AMA disclaimer

For questions regarding the use of CPT codes, contact the AMA CPT Information and Education Services at 800.634.6922 or via the Internet at http://www.ama-assn.org.

SNOMED CT[15]

Systematized Nomenclature of Medicine Clinical Terms or SNOMED CT. is a registered trademark of SNOMED International. SNOMED CT contains over 357,000 health care concepts with unique meanings and formal logic-based definitions organized into hierarchies. The fully populated table with unique descriptions for each concept contains more than 957,000 descriptions. Approximately 1.37 million semantic relationships exist to enable reliability and consistency of data retrieval.

SNOMED International maintains the SNOMED CT technical design, the core content architecture, and the SNOMED CT Core content. SNOMED CT Core content includes the technical specification of SNOMED CT and fully integrated multi-specialty clinical content. The Core content includes the concepts table, description table, relationships table, history table, and ICD-9-CM mapping, and the Technical Reference Guide.

Each SNOMED record corresponds to a single observation. The SNOMED codes are not intended to transmit all possible information about an observation, or procedure. They are only intended to identify the observation or procedure. The SNOMED code for a name is unique and permanent.

SNOMED CT combines the content and structure of the SNOMED Reference Terminology (SNOMED RT) with the United Kingdom's Clinical Terms Version 3 (formerly known as the Read Codes).

---

[14] American Medical Association/Centers for Medicare & Medicaid Services. *Current Procedural Terminology (CPT) Copyright Notices and Disclaimers and Point and Click License*. 2011.

[15] Centers for Medicare & Medicaid Services. *Data Submission Specifications Utilizing HL7 QRDA Implementation Guide Based on HL7 CDA Release 2.0 Version: 2.0* Last Modified: July 01, 2010 Available at: https://www.cms.gov/PQRI/20_AlternativeReportingMechanisms.asp#TopOfPage Accessed August 7, 2012.

For information on obtaining the standard, contact:

SNOMED International
College of American Pathologists
325 Waukegan Rd
Northfield, IL 60093-2750
http://www.ihtsdo.org

Logical Observation Identifier Names and Codes (LOINC®)[16]

LOINC codes are available for commercial use without charge, subject to the terms of a license that assures the integrity and ownership of the codes. The LOINC database provides sets of universal names and ID codes for identifying laboratory and clinical observations and other units of information meaningful in cancer registry records.

Each LOINC record corresponds to a single observation. The LOINC codes are not intended to transmit all possible information about a test or observation. They are only intended to identify the observations. The LOINC code for a name is unique and permanent. LOINC codes must always be transmitted with a hyphen before the check digit (e.g., "10154-3"). The numeric code is transmitted as a variable length number, without leading zeros.

LOINC codes are copyrighted by Regenstrief Institute and the Logical Observation Identifier Names and Codes Consortium.

The LOINC database can be obtained from:

Regenstrief Institute
1001 West 10th Street RG-5
Indianapolis, IN 46202

RxNorm[17]

RxNorm is the recommended national standard for medication vocabulary for clinical drugs and drug delivery devices produced by the National Library of Medicine (NML). RxNorm is intended to cover all prescription medications approved for human use in the United States.

Because every drug information system that is commercially available today follows somewhat different naming conventions, a standardized nomenclature is needed for the smooth exchange of information. The goal of RxNorm is to allow various systems using different drug nomenclatures to share data efficiently at the appropriate level of abstraction.

---

[16]Centers for Medicare & Medicaid Services. *Data Submission Specifications Utilizing HL7 QRDA Implementation Guide Based on HL7 CDA Release 2.0 Version: 2.0* Last Modified: July 01, 2010 Available at: https://www.cms.gov/PQRI/20_AlternativeReportingMechanisms.asp#TopOfPage Accessed August 7, 2012.

[17]Centers for Medicare & Medicaid Services. *Data Submission Specifications Utilizing HL7 QRDA Implementation Guide Based on HL7 CDA Release 2.0 Version: 2.0* Last Modified: July 01, 2010 Available at: https://www.cms.gov/PQRI/20_AlternativeReportingMechanisms.asp#TopOfPage, Accessed August 7, 2012

Each (RxNorm) clinical drug name reflects the active ingredients, strengths, and dose form comprising that drug. When any of these elements vary, a new RxNorm drug named is created as a separate concept.

More information can be found at http://www.nlm.nih.gov/research/umls/rxnorm/docs/index.html.

9. eMeasure Specifications

9.1 Introduction

The purpose of this section is to guide measure developers on how to develop and document eMeasure specifications for either an adapted (retooled) measure or a new (de novo) measure. Final technical specifications should be detailed and concise, and provide the comprehensive details that allow the measure to be collected and implemented consistently, reliably, and effectively. The process of developing eMeasure specifications occurs throughout the measure development process. In the information gathering stage, the measure contractor identifies whether existing measures may be adapted/retooled to fit the desired purpose. If no measures are identified that match the desired purpose for the measure, the contractor must work with its technical expert panel (TEP) to develop new measures. Depending on the information gathering findings, the TEP will consider potential measures that are newly proposed or are derived from existing measures. The measure contractor submits the list of candidate measures, selected with TEP input, to the Centers for Medicare & Medicaid Services (CMS) Contracting Officer Representative/Government Task Leader (COR/GTL) for approval. Upon approval from the COR/GTL, the measure contractor proceeds with the development of detailed technical specifications for the measures.

The detailed process for development of a de novo measure is outlined in the preceding Blueprint sections. Contractors are expected to follow these procedures, which often include special considerations for eMeasures. Where the process deviates from the way other types of measures are developed, it is denoted with the following icon (scaled to fit the text):

Figure 9-1 eMeasure icon denoting special considerations for eMeasures in the Blueprint.

Developing eMeasure specifications is an iterative process. A brief overview of this process is as follows.

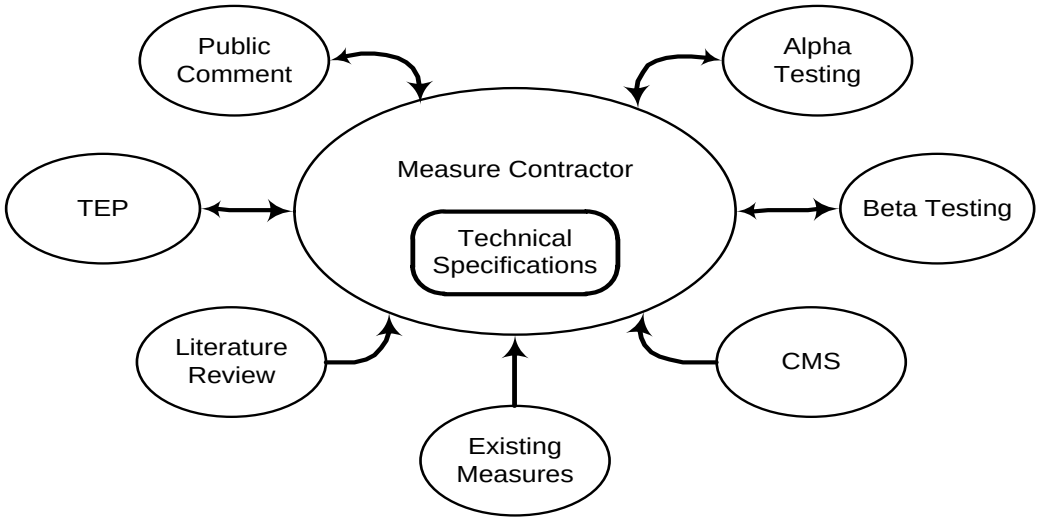
1. Prior to drafting initial eMeasure specifications, the measure contractor should consider the data elements necessary for the proposed measure, and conduct preliminary feasibility assessments to confirm availability of the information within a structured format within the electronic health record (EHR). This will better ensure that a developed measure passes feasibility assessments during beta (field) testing. (Refer to the Measure Testing section.)
2. The measure contractor then drafts the initial specifications. The TEP will review and advise the measure contractor of the recommended changes. (Refer to the Technical Expert Panel section.)
3. If directed by the COR/GTL, public comments may be obtained regarding the draft measures. (Refer to the Public Comment section.) Comments received during the public comment period

will be reviewed and taken into consideration by the measure contractor, CMS, and the TEP. This often results in revisions to the measure specifications.

4. During the development process, alpha (formative) testing of the measure occurs. (Refer to the Measure Testing section.)
5. When the specifications are more fully developed, beta (field) testing occurs. (Refer to the Measure Testing section.) As a result, technical specifications will continue to evolve and become more detailed and precise.
6. After measure testing is complete, obtain additional public comments using the formal process outlined in the Public Comment section. All of these factors will enhance the technical specifications and make the measure more valid and reliable.

Figure 9-2 illustrates the inputs to the process

Figure 9-2 Factors Influencing the Development of the Measure Technical Specifications



eMeasures, derived from Clinical Quality Measures for electronic capture, will be authored in the National Quality Forum (NQF)-developed Measure Authoring Tool (MAT). This section is a conceptual guide to eMeasure specifications development, intended to be used in conjunction with the tool and the tool's user guide. Consistency in the representation of all eMeasures used in CMS programs requires that measure developers follow the guidance in this document and adhere to the tool's user guide. The Measure Authoring Tool’s output is designed so that the contractor may submit the completed eMeasure to NQF for endorsement if desired.

As further described in this section, eMeasures are written to conform to the Health Quality Measures Format standard (HQMF), published in 2010 as a Health Level Seven (HL7) Draft Standard for Trial Use (DSTU). A DSTU is a draft standard, issued at a point in the standards development lifecycle when many but not all of the guiding requirements have been clarified. A DSTU is intended to be tested and

ultimately taken back through the HL7 ballot process, to be formalized into an American National Standards Institute (ANSI)-accredited standard.

To ensure consistency in the eMeasure development process, CMS convenes an ongoing eMeasures Issues Group (eMIG) meeting where new requirements can be vetted. The eMIG serves as a forum to discuss and propose solutions for issues encountered during the development of eMeasures. A fundamental activity of the eMIG is to develop additional guidance for issues encountered by measure developers. Solutions proposed and presented at the eMIG meetings are expected to address areas where common standards or approaches have not yet been specified, or where the *Blueprint for the CMS Measures Management System* does not supply guidance. This ever-expanding body of information discussed in the eMIG will be incorporated into the Blueprint guidance provided to developers. Measure developers should contact a member of the CMS Measures Management System team at eMIG@hsag.com for inclusion in these regularly scheduled eMIG meetings.

CMS anticipates that a final ANSI-accredited eMeasure standard may represent certain constructs differently than the way they are represented through the NQF-developed Measure Authoring Tool or the supplemental CMS eMeasure editorial guidelines. Such differences will be published in parallel with the ANSI standard. For now, measure developers need to be knowledgeable of this section, the Measure Authoring Tool's user guide, and the eMIG recommendations. When in doubt, measure developers should consult with their COR/GTL. Given the evolving nature of Health Information Technology (HIT) standards, this Blueprint section will be reviewed on an ongoing basis and updated quarterly as changing standards necessitate.

A fully constructed eMeasure is an XML document. It can be viewed in a standard web browser, when associated with an eMeasure rendering style sheet supplied by CMS. Exports of an eMeasure from the MAT include the eMeasure XML file, eMeasure HTML file, eMeasure style sheet and a standalone spreadsheet that contains all of the value sets referenced by the measure. Note that while value sets are authored in the Measure Authoring Tool, they are not part of the published eMeasure.

9.2 Deliverables

- eMeasure XML file
- eMeasure human-readable rendition (HTML) file
- eMeasure style sheet
- Value sets—available through the National Library of Medicine [NLM] Value Set Authority Center [VSAC] https://vsac.nlm.nih.gov/ for eMeasures in use in the EHR Incentive Program
- Measure Justification form (required only for de novo eMeasures and located in the Measure Development section)

9.3 Health Quality Measures Format (HQMF)

A clinical quality measure encoded in the Health Quality Measures Format is referred to as an “eMeasure” or “eCQM.” eMeasure is an HL7 standard for representing a health quality measure as an electronic XML document. Through standardization of a measure’s structure, metadata, definitions,

and logic, the HQMF provides for quality measure consistency and unambiguous interpretation. HQMF is a component of a larger quality end-to-end framework in which providers will ideally be able to push a button and import these eMeasures into their Electronic Health Records (EHRs). The eMeasures can be turned into queries that automatically gather data from the EHR's data repositories and generate reports for quality reporting. From there, individual and/or aggregate patient quality data can be transmitted to the appropriate agency.

Major components of an HQMF document include a Header and a Body. The Header identifies and classifies the document and provides important metadata about the measure. The HQMF Body contains eMeasure sections, e.g., data criteria, population criteria, and supplemental data elements. Each section can contain narrative descriptions and formally encoded HQMF entries.

To aid in the creation of an eMeasure, NQF has developed a Measure Authoring Tool.[1] The authoring tool uses a graphical user interface (GUI) to guide measure developers through the measure authoring process to create an eMeasure. The tool hides much of the complexity of the underlying HQMF from the developer. This section of the Blueprint assumes that measure developers will be using the authoring tool, and describes the eMeasure authoring process from that perspective. When seeking NQF endorsement for an eMeasure, it is important to note that the preferred submission format is based on MAT output.

eMeasure is one component of a larger quality framework. When measures are unambiguously represented as eMeasures, they can be used to guide collection of EHR and other data, which is then assembled into quality reports and submitted to organizations such as CMS. The transmission format (e.g., the interoperability specification that governs how the individual or aggregate patient data is to be communicated to CMS) is another important component of the quality framework. Developing the transmission format along with an eMeasure maximizes internal consistency.

9.4 National Quality Forum Quality Data Model

The Health Information Technology Expert Panel (HITEP), a committee of content experts convened by NQF, created the Quality Data Model (QDM, formerly referred to as Quality Data Set or QDS).[2] The QDM is an information model that defines concepts that recur across quality measures and clinical care and is intended to enable automation of EHR use. The QDM contains six components: (1) QDM element—an atomic unit of information that has precise meaning to communicate the data required within a quality measure), (2) category—a particular group of information that can be addressed in a quality measure, (3) datatype—context expected for any given QDM element, (4) code set—standard vocabulary or other classification system to define a QDM element's category, (5) value set—used to define an instance of a category , and (6) attribute—specific detail about a QDM element. A QDM element is specified by selecting a category, the datatype which the category is expected to be found with respect to electronic clinical data, a value set from an appropriate taxonomy (or vocabulary) , and all required attributes. For example, defining a value set for diabetes and applying the category,

[1] https://mat.qualityforum.org/Login.html
[2] http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx

diagnosis, and the datatype “diagnosis, active” forms the QDM element, diagnosis, acitve: diabetes, as a specific instance for use in a measure.

9.5 Building Block Approach to eMeasures

NQF has developed a building-block approach to develop eMeasures. This approach, built into the NQF-developed authoring tool, takes each datatype (e.g., diagnosis, active) in the QDM and represents it as a reusable pattern. Coupled with a value set (e.g., SNOMED CT, ICD 10 CM and ICD9 CM, codes for pneumonia), a quality pattern becomes a QDM element representing an HQMF data criterion (e.g., “active diagnosis of pneumonia”). Data criteria are assembled (using Boolean and other logical, temporal, and numeric operators) into population criteria (e.g., “Denominator = active diagnosis of pneumonia, AND age > 18 at time of hospital admission”), thereby creating a formal representation of a quality measure that can be processed by a computer.

Thus, at a high level, the process of creating an eMeasure is to map measure data elements to the correct datatypes in the QDM, associate each datatype with the correct value set(s) to create data criteria, and then assemble the data criteria into population criteria.

9.6 Measure Authoring Tool

As described above, NQF has developed a web-based Measure Authoring Tool, a software authoring tool that measure developers use to create eMeasures. The authoring tool allows measure developers to create their eMeasures in a highly structured format using the QDM and healthcare industry standard vocabularies.[3]

The QDM 2.1.1.1 (October 2012), is the current version released in the MAT, which is also the version cited under Meaningful Use Stage 2 (MU2).

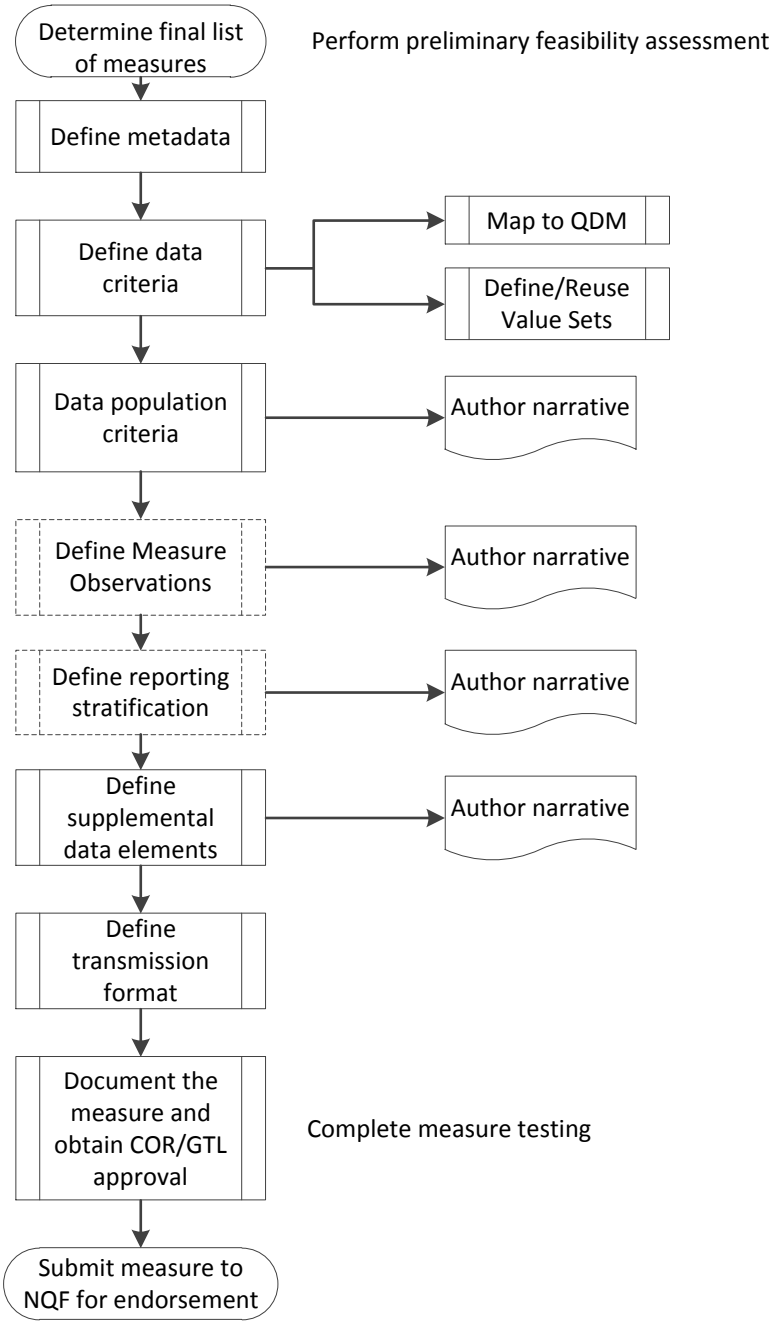
The Measure Authoring Tool does not require measure developers to have an extensive knowledge of the HQMF standard. It is based on the QDM and the building-block approach to creating eMeasures. It supports common-use cases and the existing patterns in the pattern library. The Measure Authoring Tool will require ongoing maintenance and support to meet any future measure authoring needs. For instance, if a measure requires a new datatype and therefore a new corresponding quality pattern, the pattern must first be developed and then added to the Measure Authoring Tool.

9.7 Procedure

When retooling or developing a new eMeasure, the developer should use the process outlined in the following figure and detailed in the sections that follow. Dashed boxes are only required under certain circumstances.

[3] http://www.qualityforum.org/Projects/e-g/eMeasure_Format_Review/eMeasure_Format_Review.aspx

Figure 9-3 eMeasure Specifications and Transmission Format Development Process



Step 1: Determine final list of measures (and perform preliminary feasibility assessment)

Based on the environmental scan, measure gap analysis, and other information gathering activities, the measure contractor submits the list of candidate measures—selected with TEP input—to the COR/GTL

for approval. These measures may be retooled or de novo. Before deciding to retool a measure, consider the following issues.

- If the existing measure is NQF-endorsed, are the changes to the measure significant enough to require resubmission to NQF for endorsement?
- Will the measure owner be agreeable to the changes in the measure specifications that will meet the needs of the current project?
- If a measure is copyright protected, are there issues relating to the measure's copyright that need to be considered?

These considerations must be discussed with the COR/GTL and the measure owner, while NQF endorsement status may need to be discussed with NQF. Upon approval from the COR/GTL, the measure contractor proceeds with the development of detailed technical specifications for the measures using the NQF-developed MAT.

Prior to drafting initial eMeasure specifications, the measure contractor should consider the data elements necessary for the proposed measure, and conduct preliminary feasibility assessments (alpha testing) to confirm availability of the information within a structured format within the electronic health record. This will better ensure that a developed measure passes feasibility assessments during beta (field) testing. (Refer to the Measure Testing section).

Step 2: Define metadata

The Header of an eMeasure document identifies and classifies the document and provides important metadata about the measure. eMeasure metadata are summarized in Table 9-1 listing elements in the order that they are conventionally displayed.

The eMeasure Header should include the appropriate information for each element as described in the *Definition* column of Table 9-1. The default for each element in the Measure Authoring Tool is a blank field, but all header fields need to have an entry. The *Preferred Term* column indicates how the developer should complete entry into the MAT. "**Required**" means that the measure developer must populate the metadata element field as defined in column 2. All eMeasure header fields must have information completed OR placement of a "**None"** or "**Not Applicable**" in the header field. Conventions for when to use "**None**" versus "**Not applicable**" have been described for each metadata element and should be entered according to *Preferred Term* column instructions (e.g., measures not endorsed by NQF should populate the metadata element NQF Number with "**Not Applicable**").

Table 9-1 eMeasure Metadata

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
eMeasure Title	The title of the quality eMeasure.		**Required**
eMeasure Identifier (Measure Authoring Tool)	Specifies the eMeasure identifier generated by the NQF-developed Measure Authoring Tool	Field is autopopulated by the Measure Authoring Tool.	**Required**
GUID	Represents the globally unique measure identifier for a particular quality eMeasure.	Field is autopopulated by the Measure Authoring Tool.	**Required**
eMeasure Version Number	A positive integer value used to indicate the version of the eMeasure.	Displays the integer (whole number) the measure developer enters. The version number is a whole integer that should be increased each time the developer makes a change to the eMeasure that in the opinion of the developer effects the substantive content, or intent of the measure OR the logic or coding of the measure . The following types of changes do not require a version number integer increase unless the developer elects to change the version number for other reasons. Typos or word changes that do not affect the substantive content or intent of the measuree.g., changing from an abbreviation for a word to spelling out the word for a test value unit of measurement, would not require an increment in version number.	**Required**
NQF Number	Specifies the NQF number	*"Optional" field in MAT* eMeasures endorsed by NQF should enter this as a 4-digit number (including leading zeros). Only include an NQF number if the eMeasure is endorsed.	**Not Applicable**

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Measurement Period	The time period for which the eMeasure applies.	MM/DD/20xx – MM/DD/20xx	**Required**
Measure Steward	The organization responsible for the continued maintenance of the eMeasure.	Can be the same as the Measure Developer.	**Required**
Measure Developer	The organization that developed the eMeasure.		**Required**
Endorsed By	The organization that has endorsed the eMeasure through a consensus-based process.	Provided by the measure steward; all endorsing organizations should be included (not specific to just NQF).	**None**
Description	A general description of the eMeasure intent	A brief narrative description of the eMeasure, such as "Ischemic stroke patients with atrial fibrillation/flutter who are prescribed anticoagulation therapy at hospital discharge."	**Required**
Copyright	Identifies the organization(s) who own the intellectual property represented by the eMeasure.	The owner of the eMeasure has the exclusive right to print, distribute, and copy the work. Permission must be obtained by anyone else to reuse the work in these ways. May also include copyright permissions (e.g., "©2010 American Medical Association. All Rights Reserved").	**None**
Disclaimer	Disclaimer information for the eMeasure.	This should be brief.	**None**
Measure Scoring	Indicates how the calculation is performed for the eMeasure (e.g., proportion, continuous variable, ratio)		**Required**

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Measure Type	Indicates whether the eMeasure is used to examine a process or an outcome over time (e.g., Structure, Process, Outcome).		**Required**
Stratification	Describes the strata for which the measure is to be evaluated. There are three examples of reasons for stratification based on existing work. These include: (1) evaluate the measure based on different age groupings within the population described in the measure (e.g., evaluate the whole <age 14-25> and each sub-stratum <14-19> and <20-25>); (2) evaluate the eMeasure based on either a specific condition, a specific discharge location, or both; (3) evaluate the eMeasure based on different locations within a facility (e.g., evaluate the overall rate for all intensive care units and also some strata include additional findings <specific birth weights for neonatal intensive care units>)		**None**
Risk Adjustment	The method of adjusting for clinical severity and conditions present at the start of care that can influence patient outcomes for making valid comparisons of outcome measures across providers. Indicates whether an eMeasure is subject to the statistical process for reducing, removing, or clarifying the influences of confounding factors to allow more useful comparisons.	Brief with instructions where complete risk adjustment methodlogy may be obtained	**None**

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Rate Aggregation	Describes how to combine information calculated based on logic in each of several populations into one summarized result. It can also be used to describe how to risk adjust the data based on supplemental data elements described in the eMeasure. (e.g., pneumonia hospital measures antibiotic selection in the ICU versus non-ICU and then the roll-up of the two).		**None**
Rationale	Succinct statement of the need for the measure. Usually includes statements pertaining to Importance criterion: impact, gap in care and evidence.		**Required**
Clinical Recommendation Statement	Summary of relevant clinical guidelines or other clinical recommendations supporting this eMeasure.		**Required**
Improvement Notation	Information on whether an increase or decrease in score is the preferred result (e.g., a higher score indicates better quality OR a lower score indicates better quality OR quality is within a range).		**None**
Measurement Duration	Field will not be used	Does not show in the style sheet.	**None**
Reference(s)	Identifies bibliographic citations or references to clinical practice guidelines, sources of evidence, or other relevant materials supporting the intent and rationale of the eMeasure.		**None**
Definition	Description of individual terms, provided as needed.	*Note—this field may be removed in the future.	**None**

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Guidance	Used to allow measure developers to provide additional guidance for implementers to understand greater specificity than could be provided in the logic for data criteria.		**None**
Transmission Format	Can be a URL or hyperlinks that link to the transmission formats that are specified for a particular reporting program.	For example, it could be a hyperlink or URL that points to the Quality Reporting Document Architecture (QRDA) Category I implementation guide, or a URL or hyperlink that points to the PQRI Registry XML specification. This is a free text field. *Further guidance forthcoming.	**None**
Initial Patient Population	The *initial patient population* refers to all patients to be evaluated by a specific performance eMeasure who share a common set of specified characteristics within a specific measurement set to which a given measure belongs. Details often include information based upon specific age groups, diagnoses, diagnostic and procedure codes, and enrollment periods.	Must be consistent with the computer-generated narrative below. The computer generated narrative is standardized and concise, and can lack the richness of full text that sometimes helps in the understanding of an eMeasure. This is especially true for eMeasures that have complex criteria, where the computer generated text may not be able to express the exact description that a measure developer would like to convey. As part of the quality assurance step, it is important to compare the manually authored narrative against the automatically rendered narrative for any discrepancies. This field will be the primary field to fully define the comprehensive eligible population for proportion/ratio eMeasures or the eligible measure population for continuous variable eMeasures.	**Required**

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Denominator	It can be the same as the initial patient population or a subset of the initial patient population to further constrain the population for the purpose of the eMeasure. Different measures within an eMeasure set may have different Denominators. Continuous Variable eMeasures do not have a Denominator, but instead define a Measure Population.	For proportion/ratio measures, include the text "Equals Initial Patient Population" where applicable.	**Not Applicable** (for continuous variable eMeasures)
Denominator Exclusions	Patients who should be removed from the eMeasure population and denominator before determining if numerator criteria are met. Denominator exclusions are used in proportion and ratio measures to help narrow the denominator. (e.g., Patients with bilateral lower extremity amputations would be listed as a denominator exclusion for a measure requiring foot exams.)		**None** (for proportion or ratio eMeasures) **Not Applicable** (for continuous variable eMeasures)
Numerator	Numerators are **used in proportion and ratio eMeasures**. In proportion measures the numerator criteria are the processes or outcomes expected for each patient, procedure, or other unit of measurement defined in the denominator. In ratio measures the numerator is related, but not directly derived from the denominator (e.g., a numerator listing the number of central line blood stream infections and a denominator indicating the days per thousand of central line usage in a specific time period).		**Not Applicable** (for continuous variable eMeasures)

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Numerator Exclusions	Numerator Exclusions are **used only in ratio eMeasures** to define instances that should not be included in the numerator data. (e.g., if the number of central line blood stream infections per 1000 catheter days were to exclude infections with a specific bacterium, that bacterium would be listed as a numerator exclusion.)		**None** (for ratio eMeasures) **Not Applicable** (for continuous variable or proportion eMeasures)
Denominator Exceptions	Denominator exceptions are those conditions that should remove a patient, procedure or unit of measurement from the denominator of the performance rate only if the numerator criteria are not met. Denominator exceptions allow for adjustment of the calculated score for those providers with higher risk populations. Denominator exceptions are **used only in proportion eMeasures**. They are not appropriate for ratio or continuous variable eMeasures. Denominator exceptions allow for the exercise of clinical judgment and should be specifically defined where capturing the information in a structured manner fits the clinical workflow. Generic denominator exception reasons used in proportion eMeasures fall into three general categories: ◆ Medical reasons ◆ Patient reasons ◆ System reasons	Be specific for medical reasons.	**None** (for proportion eMeasures) **Not Applicable** (for ratio or continuous variable Measures)

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Measure Population	Measure population **is used only in continuous variable eMeasures**. It is a narrative description of the eMeasure population. (e.g., all patients seen in the Emergency Department during the measurement period).	For continuous variable eMeasures, include the text "Equals All in Initial *Patient* Population." Then add any specific additional criteria if needed.	**Not Applicable** (for ratio or proportion eMeasures)
Measure Observations	Measure observations are **used only in continuous variable eMeasures.** They provide the description of how to evaluate performance, (e.g., the mean time across all Emergency Department visits during the measurement period from arrival to departure). Measure observations are generally described using a statistical methodology such as: count, median, mean, etc.		**Not Applicable** (for ratio or proportion eMeasures)
Supplemental Data Elements	CMS defines four required Supplemental Data Elements (payer, ethnicity, race, and sex), which are variables used to aggregate data into various subgroups. Comparison of results across strata can be used to show where disparities exist or where there is a need to expose differences in results. Additional supplemental data elements required for risk adjustment or other purposes of data aggregation can be included in the Supplemental Data Element section.	Due to the four CMS required fields, the Developer must always populate with payer, ethnicity, race and sex. For CMS measures use the following language in Supplemental Data section (January 2012 release of MAT) "For every patient evaluated by this measure also identify payer, race, ethnicity and sex." Other information may be added for other measures.	**Required**

Conventions for developing eMeasures

Developers should use the following conventions in Table 9-2 when developing the CMS eMeasures to ensure standardization of the measures for display of specifications and quality reporting:

Table 9-2 Conventions for Developing eMeasures

Data Element or Item	Convention
Calculation of Age for Ambulatory Measures	To be included in the eMeasure, the patient's age must be ≥ IPP inclusion criterion before the start of the Measurement Period. The Measurement Period is January 1-December 31, 20xx. e.g., If the inclusion criterion for the measure is ≥18 years of age, then the patient must reach the age of 18 prior to the beginning of the measurement year.
Calculation of Age for Hospital Measures	In the context of hospital measures, patient age can be defined in multiple ways, all of which retain significance in the context of a single episode of care. While the most commonly used age calculation is patient age at admission (the age of the patient (in years) on the day the patient is admitted for inpatient care), there are situations where specific population characteristics or the measure's intent require distinct calculations of patient age: ◆ Patient age at admission (in years) [refer to example in note A below] = Admission date – Birth date ◆ Newborn patient age at admission (in days) [refer to example in note B below] = Admission date - Birth date ◆ Patient age at discharge (in years) [refer to example in note C below] = Discharge date - Birth date ◆ Patient age at time of event (in years) [refer to example in note D below] = Event date - Birth date **Note A**—An example of a corresponding representation in HQMF for an IPP inclusion criterion for patient age range: ◆ "Patient Characteristic Birthdate: birth date" <= 65 years starts before start of x; AND ◆ "Patient Characteristic Birthdate: birth date" >= 18 years starts before start of x **Note B**—An example of a corresponding representation in HQMF: Patient is <= 20 days old at time of admission: ◆ "Patient Characteristic Birthdate: birth date" <= 20 days starts before start of "Encounter Performed: hospital encounter (admission)" **Note C**—An example of a corresponding representation in HQMF: Patient is >= 10 years old at time of discharge: ◆ "Patient Characteristic Birthdate: birth date" >= 10 years starts before start of "Encounter Performed: hospital encounter (discharge)" **Note D**—An example of corresponding representation in HQMF: Patient is <= 21 at time of hepatitis vaccination: ◆ "Patient Characteristic Birthdate: birth date" <= 21 years starts before start of "Medication Administered: hepatitis vaccine (date time)"

Step 3: Define data criteria and key terms

Map to QDM

As previously indicated, the process of creating a new eMeasure is to map measure data elements to the correct datatypes in the QDM, associate each datatype with the correct value set(s) to create data criteria, and then assemble the data criteria into population criteria. The process works somewhat differently for retooled measures.

Retooled measures

Measure developers need to identify data elements in the existing paper-based measure and map them to QDM datatype to define data criteria in a quality measure retooling scenario (e.g., when transforming an existing paper measure into an eMeasure). Developers will associate each QDM datatype with an existing value set or create a new value set if one does not exist. The HIT Standards Committee (HITSC) has developed a set of recommendations to report quality measure data using clinical vocabulary standards. Recognizing that immediate use of clinical vocabularies may be a challenge, the HITSC also developed a transition plan that includes a list of acceptable transition vocabularies and associated timeframes for use. It is important to note that the recommendations do not apply beyond the domain of eMeasure development and maintenance. Value sets should be aligned with HIT Standards Committee (HITSC) recommended vocabulary standards, and should include the transitional vocabularies: ICD-10-CM, ICD-10-PCS, ICD-9-CM, Current Procedural Terminology, CPT®, and HCPCS where applicable.

New measures

Measure developers are expected to author a new measure directly in eMeasure format using the Measure Authoring Tool. A data criterion will be constructed based on a QDM datatype. Developers will associate each QDM datatype with an existing value set or create a new value set if one does not exist, and define additional attributes if applicable.

Clearly define any time windows.

Time windows must be stated whenever they are used to determine cases for inclusion in the denominator, numerator, or exclusions. Any index event used to determine the time window is to be stated. Avoid the use of ambiguous semantics when referring to time intervals.

Example:

- Medication reconciliation must be performed within **30 days following hospital discharge.** Thirty days is the time window and the hospital discharge date is the index event.
  - This example illustrates ambiguity in interpretation: “30 days” should be clearly identified as calendar days, business days, etc. within the measure guidelines to prevent unintended variances in reportable data.

Appendix 9c1 provides definitions for time units and intervals, while Appendix 9c2 provides time interval calculation conventions. These conventions are intended to standardize the time calculation

units for durations e.g., difference between two date/time elements, typically with time relationships defined in the QDM, such as "starts after start of" and "ends before start of".

Define/reuse Value Sets

Value sets are specified and bound to coded data elements in an eMeasure. When creating value sets, it is important to align with the vocabulary recommendations made by HITSC Clinical Quality Workgroup and Vocabulary Task Force of The Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services (HHS). The HITSC recommended vocabulary standards are listed in Tables 9-3 through 9-5 below. Table 9-5 illustrates the vocabularies as they relate to the clinical concepts of the QDM. Measures that are retooled or developed must include ICD-10 codes where applicable.

On January 16, 2009, the U.S. Department of Health and Human Services (HHS) released a final rule mandating that everyone covered by the Health Insurance Portability and Accountability Act (HIPAA) must implement ICD-10 for medical coding by October 1, 2013. However, on April 17, 2012 HHS published a proposed rule that would delay the compliance date for ICD-10 from October 1, 2013 to October 1, 2014.[4]

Table 9-3 ONC HIT Standards Committee Recommended Vocabulary Standards Summary

Clinical Vocabulary Standards:	Others:
SNOMED CT	CVX
LOINC	CDC-PHIN/VADS
RxNorm	UCUM
	ISO-639

Table 9-4 ONC HIT Standards Committee Transition Vocabulary Standards Summary and Plan

Transition Vocabulary	Transition period: Acceptable for reporting eMeasure results for	Final date for reporting eMeasure results
ICD-9-CM	Dates of service before the implementation of ICD-10	Not acceptable for reporting eMeasure results for services provided after the implementation of ICD-10
ICD-10-CM	Dates of service on or after the implementation of ICD-10	Final Date: one year after MU-3 is effective

[4]Centers for Medicare & Medicaid Services. *ICD-10, Statute and Regulations.* Available at: http://www.cms.gov/ICD10/02d_CMS_ICD-10_Industry_Email_Updates.asp#TopOfPage. Accessed July 31, 2012.

Transition Vocabulary	Transition period: Acceptable for reporting eMeasure results for	Final date for reporting eMeasure results
ICD-10-PCS	Dates of service on or after the implementation of ICD-10	Final Date: one year after MU-3 is effective
CPT	Acceptable during MU 1,2,3 if unable to report using clinical vocabulary standards	Final Date: one year after MU-3 is effective
HCPCS	Acceptable during MU 1,2,3 if unable to report using clinical vocabulary standards	Final Date: one year after MU-3 is effective

When specifying eMeasures using value sets, measure contractors should note the following:

- Generic drug names—Value sets will contain generic, rather than brand drug names
- ICD9CM and ICD10CM Group Codes—Codes that are not valid for clinical coding should not be included in value sets. Specifically, codes that are associated with sections or groups of codes should not be used in value sets. Examples are provided below.
  - Use the following fifth-digit sub-classification with category 948 to indicate the percent of body surface with third degree burn:
    - 0—less than 10 percent or unspecified
    - 1—10-19 percent
    - 2—20-29 percent
    - 3—30-39 percent
    - 4—40-49 percent
    - 5—50-59 percent
    - 6—60-69 percent
    - 7—70-79 percent
    - 8—80-89 percent
    - 9—90 percent or more of body surface
  - 632 Missed abortion—This is a stand-alone code—does not require any additional digits to be valid.
  - 490 Bronchitis (diffuse) (hypostatic) (infectious) (inflammatory) (simple)—This is a stand-alone code and does not have any other association for it. This is a 3-digit billable code.
  - 633 Ectopic pregnancy—This is a non-billable code—this must have additional digits to be valid (e.g., 633.1 Tubal pregnancy).

Table 9-5 Quality Data Model Categories with ONC HIT Standards Committee Recommended Vocabularies

General Clinical Concept	Quality Data Model Category	Quality Data Model Attribute	Clinical Vocabulary Standards	Transition Vocabulary
Adverse Effect/ Allergy/ Intolerance	*Device*	*Reaction*	SNOMED CT	N/A
	Diagnostic Study	*Reaction*	SNOMED CT	

General Clinical Concept	Quality Data Model Category	Quality Data Model Attribute	Clinical Vocabulary Standards	Transition Vocabulary
	Intervention	*Reaction*	SNOMED CT	
	Laboratory Test	*Reaction*	SNOMED CT	
	Medication	*Reaction*	RxNorm	
	Procedure	*Reaction*	SNOMED CT	
	Substance	*Reaction*	SNOMED CT	
Substance	*Substance*	*N/A*	SNOMED CT	N/A
Condition/Diagnosis/ Problem	*Condition/Diagnosis/ Problem*	N/A	SNOMED CT	ICD-9-CM, ICD-10-CM
Symptom	*Symptom*	N/A	SNOMED CT	N/A
Encounter (any patient-provider interaction, e.g. phone call, email regardless of reimbursement status, status— includes traditional face-to-face encounters)	*Encounter*	*N/A*	SNOMED CT	CPT, HCPCS, ICD-9-CM Procedures, ICD-10-PCS
Device	*Device*	*N/A*	SNOMED CT	N/A
Physical exam finding	*Physical Exam*	*result*	SNOMED CT	N/A
Laboratory test (names)	*Laboratory Test*	*N/A*	LOINC	N/A
Laboratory test (results)	*Laboratory Test*	*result*	SNOMED CT	N/A
Diagnostic study test names	*Diagnostic Study*	*N/A*	LOINC	HCPCS
Diagnostic study test results	*Diagnostic Study*	*result*	SNOMED CT	N/A
Units of Measure for results	N/A	N/A	UCUM-(The Unified Code for Units of Measure)	N/A
Intervention	*Intervention*	*N/A*	SNOMED CT	CPT, HCPCS, ICD-9-CM Procedures, ICD-10-PCS

General Clinical Concept	Quality Data Model Category	Quality Data Model Attribute	Clinical Vocabulary Standards	Transition Vocabulary
Procedure	*Procedure*	*N/A*	SNOMED CT	CPT, HCPCS, ICD-9-CM Procedures, ICD-10-PCS
Patient characteristic, preference, experience (expected answers for questions related to patient characteristic, preference, experience)	*Applies to all QDM Categories Except System Characteristic*	*Patient preference*	SNOMED CT	N/A
Functional Status	*Functional status*	*N/A*	SNOMED CT	N/A
Categories of function	*Functional status*	*result*	ICF-(International Classification of Functioning, Disability, and Health)	N/A
Communication	*Communication*	*N/A*	SNOMED CT	CPT, HCPCS
Assessment instrument questions (questions for patient preference, experience, characteristics)	*Individual Characteristic*	*N/A*	LOINC	N/A
Medications (administered, excluding vaccines)	*Medication*	*N/A*	RxNorm	N/A
Vaccines (administered)	*Medication*	*N/A*	CVX	N/A
Patient characteristic (Administrative Gender, DOB)	*Individual Characteristic*	*N/A*	CDC-Public Health Information Network (PHIN)/Vocabulary Access and Distribution System (VADS) http://www.cdc.gov/phin/activities/vocabulary.html	N/A

General Clinical Concept	Quality Data Model Category	Quality Data Model Attribute	Clinical Vocabulary Standards	Transition Vocabulary
Patient Characteristic (Ethnicity, Race)	*Individual Characteristic*	*N/A*	OMB Ethnicity/Race (scope) http://www.cdc.gov/phin/library/resources/vocabulary/CDC%20Race%20&%20Ethnicity%20Background%20and%20Purpose.pdf – expressed in PHIN VADS as value sets as the vocabulary	N/A
Patient characteristic (Preferred language)	*Individual Characteristic*	*N/A*	ISO-639-1:2002	N/A
Patient characteristic (Payer)	*Individual Characteristic*	*N/A*	Payer Typology (Public Health Data Standards Consortium Payer Typology) (scope)http://www.phdsc.org/standards/pdfs/SourceofPaymentTypologyVersion5.0.pdf In PHIN VADS (vocabulary) http://phinvads.cdc.gov/vads/ViewCodeSystemConcept.action?oid=2.16.840.1.113883.221	N/A

To the extent possible, use existing QDM value sets when developing new eMeasures. The developer should look at the existing library of value sets to determine if any exist that define the clinical concepts described in the measure. If so, these should be used, rather than creating a new value set. This promotes harmonization in addition to decreasing the time needed to research the various vocabularies to build a new list.

CMS measure contractors should refer to the periodic updates to these guidelines issued by the eMIG for the most up to date vocabulary recommendations. Other developers, not involved in eMIG, may refer to the ONC HIT Standards Committee Web site http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__health_it_standards_committee/1271.

At times there may be a need to request new SNOMED-CT concepts. The request should be submitted through the U.S. SNOMED CT Content Request System (USCRS) of National Library of Medicine (NLM).[5]

[5]U.S. SNOMED CT Content Request System. https://uscrs.nlm.nih.gov/

Measure developers must sign up for a UMLS Terminology Services (UTS) account to log into the USCRS.[6]

There may also be a need to request new LOINC concepts. Instructions and tools to request LOINC concepts can be found at the LOINC Web site http://loinc.org/submissions/new-terms.

Retooling or creating a new measure may require reusing existing value sets or defining new value sets. Measure developers should define a value set as an enumerated list of codes. For example, a diabetes mellitus value set may include an enumerated list of fully specified ICD-9-CM codes, such as 250.0, 250.1, and 250.2 and SNOMED CT and ICD-10-CM codes as well. Several tools exist to help build and maintain quality measure value sets. Some of these tools can take value set criteria (e.g., "all ICD9 codes beginning with 250*") and expand them into an enumerated list. Where such value set criteria exist, they should be included as part of the value set definition.

In order to identify the recommended vocabularies as defined by the HITSC, it may be necessary to identify multiple subsets for a measure data element (e.g., a SNOMED-CT subset, an ICD-9-CM subset, and an ICD-10-CM subset). NQF has defined a grouping mechanism for this scenario. Measure developers define separate value sets for different code systems, such as a Diabetes Mellitus SNOMED-CT subset and a Diabetes Mellitus ICD-9-CM subset. They then define a Diabetes Mellitus Grouping value set that combines the two subsets, and associate the grouped value set with the measure data element. Where possible, a measure developer should reuse existing value sets. NQF is exploring the sharing of value sets across measure developers.

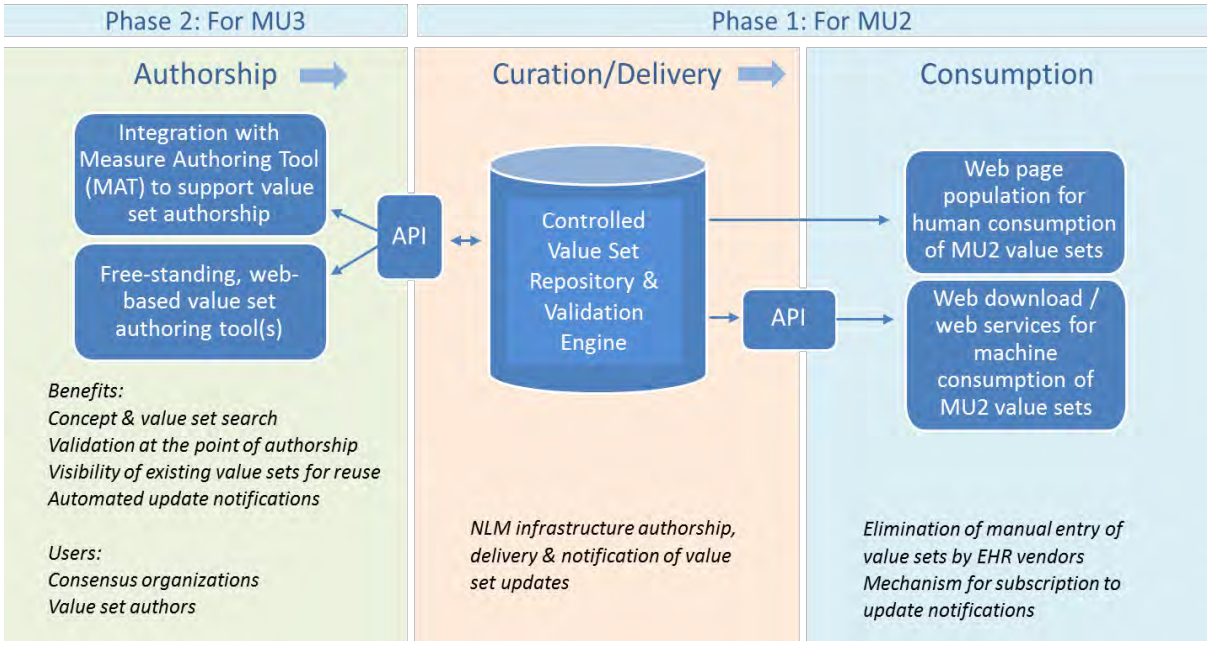
When defining a value set, measure developers may need to include codes that are no longer active in the target terminology. For example, a measure developer may need to include retired ICD-9-CM codes in a value set so that historic patient data, captured when the ICD-9-CM codes were active, also satisfies a criterion. Measure developers need to carefully consider the context in which their value sets will be used to ensure that the full list of allowable codes is included.

Note that while value sets are authored in the Measure Authoring Tool, they are not part of the published eMeasure. An eMeasure value sets spreadsheet is no longer generated as part of an eMeasure package, value sets information will be available from the online VSAC established by NLM[7]. The NLM will support ongoing maintenance of controlled value sets through this publicly available authoritative repository.

To improve value set authorship, curation, and delivery, for Meaningful Use Stage 2, NLM has performed quality assurance checks to assess the validity of value set codes, terms and associated vocabularies. In the future, NLM will develop author support tools and validation services that will be integrated into measure authoring tools to support value set development. Figure 9-4 shows the National Library of Medicine's vision for value set management.

[6] UMLS Terminology Services. https://uts.nlm.nih.gov/home.html
[7] Value Set Authority Center, National Library of Medicine. http://vsac.nlm.nih.gov

Figure 9-4 Vision for Robust Value Set Management[8]



For the value set quality assurance checks, NLM assesses the validity of code, code system, and description of each value set code, and shares the analysis result with the measure developers. Measure developers should take proper actions as specified by NLM based on the analysis outcome. If corrections are identified, measure developers should correct the value sets using the Measure Authoring Tool.

Step 4: Define population criteria

Population criteria are assembled from the underlying data criteria. The populations for a particular measure depend on the type of Measure Scoring (Proportion, Ratio, Continuous Variable), as shown in Table 9-6. The definitions for these populations are defined in Appendix 9d.

Table 9-6 Measure Populations Based on Type of Measure Scoring

	Initial Patient Population	Denominator	Denominator Exclusion	Denominator Exception	Numerator	Numerator Exclusion	Measure Population
Proportion	R	R	O	O	R	NP	NP

[8]U.S. National Library of Medicine: Value Set Validation: Author Info and Instructions.

	Initial Patient Population	Denominator	Denominator Exclusion	Denominator Exception	Numerator	Numerator Exclusion	Measure Population
Ratio	R	R	O	NP	R	O	NP
Continuous Variable	R	NP	NP	NP	NP	NP	R

R=Required. O=Option. NP=Not Permitted.

For a continuous variable measure, the population for a single measure is simply a subset of the Initial Patient Population for the measure set representing patients meeting criteria for inclusion in the measure.

However, for a proportion measure, there is a fixed mathematical relationship between population subsets. These mathematical relationships, and the precise method for calculating performance, are detailed in Appendix 9e.

Denominator exclusions vs. denominator exceptions

Definitions of each population are presented in the definitions in Table 9-1. Refer to the Technical Specifications section (Step 2) for guidance regarding when to use Denominator Exclusions versus Denominator Exceptions. There is a significant amount of discussion on the use of exclusions and exceptions, particularly the ability to capture exceptions in electronic health records. There is no agreed upon approach, however there seems to be consensus that exceptions provide valuable information for clinical decision-making. Contractors that build exceptions into measure logic should be cautioned that—once implemented—exception rates may be subject to reporting, auditing and validation of appropriateness. The difficulty in capturing exceptions as a part of clinical workflow makes the incorporation of exclusions more desirable in an EHR environment.

Assemble data criteria

Measure developers will use Boolean operators (AND and OR) to assemble data criteria to form population criteria (e.g., “Numerator = DataCriterion1 AND NOT (DataCriterion2)”). In addition to the Boolean operators, measure developers can also apply appropriate temporal context and comparators such as “*during*”, relative comparators such as “*first*” and “*last*”, EHR context (defined below), and more. Conceptual considerations are provided here. Refer to the Measure Authoring Tool user guide for authoring details.

Temporal context and temporal comparators

The QDM recommends that all elements have a date/time stamp. These date/time stamps are required by an eMeasure for any inferences of event timing (e.g., to determine whether or not

DataCriterion1 occurred before DataCriterion2; to determine if a procedure occurred during a particular encounter; etc.).

Relative timings allow a measure developer to describe timing relationships among individual QDM elements to create clauses that add meaning to the individual QDM elements. Relative timings are described in detail in the *Quality Data Model Technical Specification*[9]. For instance, "*starts before or during*" is a relative timing statement that specifies a relationship in which the source act's effective time starts before the start of the target or starts during the target's effective time. An example of this is "A pacemaker is present at any time *starts before or during* the measurement period". QDM documentation also specifies a list of functions. Functions specify sequencing (ordinality) and provide the ability to specify a calculation with respect to QDM elements and clauses containing them. It includes functions such as "FIRST", "SECOND", "LAST", and "RELATIVE FIRST." Measure developers should refer to the QDM technical specification document for descriptions and examples for a particular relative timing comparator and function.

Other data relationships

A data element in a measure can be associated with other data elements to provide more clarity. These relationships include "Is Authorized By" (used to express that a patient has provided consent); "Is Derived By" (used to indicate a result that is calculated from other values); "Has Goal Of" (used to relate a Care Goal to a procedure); "Causes" (used to relate causality); and "Has Outcome Of" (used to relate an outcome to a procedure as part of a care plan). Refer to MAT documentation for adding these relationships into an eMeasure.

Data Source

"Data Source" defines where to find the data needed to resolve a criterion. For instance, if the data source is an EHR and the context is "problem list", for a criterion such as "active problem of hypertension", then one would expect that the information regarding whether or not a patient has an active problem of hypertension would be found in the EHR's problem list. Note however that data sources in eMeasures are more suggestive than prescriptive. In some cases, particularly where Meaningful Use has not yet standardized the data source (e.g., in a criterion such as "patient refused treatment"), the data necessary to satisfy the criterion will be found in various places in various data sources. Consistent application of a standard set of data sources and contexts will lower the barrier to automated electronic reporting of quality and thus lead to consistent data storage.

Author narrative

In addition to the brief narrative description of each population's criteria that are part of the measure metadata, a narrative representation of the HQMF logic is also automatically generated by the authoring tool. The computer generated narrative is standardized and concise, and can lack the richness of full text that sometimes helps in the understanding of a measure. This is especially true for measures that have complex criteria, where the computer generated text may not be able to express

[9] http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx#t=2&s=&p=4%7C

the exact description that a measure developer would like to convey. As part of the quality assurance step, it is important to compare the manually authored narrative against the automatically rendered narrative for any discrepancies.

Step 5: Define measure observations

This step is only applicable to Continuous Variable measures. Measure Observations are required for Continuous Variable measures, but are precluded for other measures.

Measure observations are variables or calculations used to score a particular aspect of performance. For instance, a measure intends to assess the use of restraints. Population criteria for the measure include “patient is in a psychiatric inpatient setting” and “patient has been restrained.” For this population, the measure defines a measure observation of “restraint time” as the total amount of time the patient has been restrained. Measure observations are not criteria, but rather, are definitions of observations, used to score a measure.

Author narrative

It is important for measure developers to write a corresponding brief narrative description of the measure observations as part of the metadata.

Step 6: Define reporting stratification

Measure developers define Reporting Strata, which are variables on which the measure is designed to report inherently (e.g., report different rates by type of intensive care unit in a facility; stratify and report separately by age group [14-19, 20-25, and total 14-25]).

Reporting strata are optional. They can be used for proportion, ratio, or continuous variable measures. A defined value set is often a necessary component of the stratification variable so that all measures report in the same manner.

Reporting stratification vs. population criteria

A variable may appear both as a reporting stratum and a population criterion. For example, a pediatric quality measure may need data aggregated by pediatric age strata (e.g., neonatal period, adolescent period), and may also define an Initial Patient Population criterion that age be less than 18.

Author narrative

It is important for measure developers to write a corresponding brief narrative description of a measure's reporting stratification as part of the metadata.

Step 7: Define supplemental data elements

CMS defines Supplemental Data Elements which are variables used to aggregate data into various subgroups. Comparison of results across strata can be used to show where disparities exist or where there is a need to expose differences in results.

Supplemental data elements are similar to reporting stratification variables in that they allow for subgroup analysis. Whereas measure developers define reporting strata, CMS defines supplemental data elements.

CMS requires that transmission formats conveying single-patient level data must also include the following supplemental data elements:

- Sex—Should be reported as structured data, where the patient's sex is coded using a value from the ONC Administrative Sex Value Set (value set 2.16.840.1.113762.1.4.1)
- Race—Should be reported as structured data, where the patient's race is coded using a value from the CDC Race Value Set[10] (value set 2.16.840.1.114222.4.11.836).
- Ethnicity—Should be reported as structured data, where the patient's ethnicity is coded using a value from the CDC Ethnicity Value Set[11] (value set 2.16.840.1.114222.4.11.837).
- Payer—Should be reported as structured data, where the patient's payer source is coded using a code from the Payer Source of Payment Typology (value Set 2.16.840.1.114222.4.11.3591) approved by PHDSC. [http://www.phdsc.org/standards/pdfs/SourceofPaymentTypologyUsersGuideVersion4.0_final.pdf].

These supplemental data elements, along with any additional elements defined for a particular eMeasure, must be present for all CMS quality measures. The Measure Authoring Tool automatically adds the supplemental data elements section to an eMeasure, when the eMeasure is exported from the tool. The value sets referenced by the supplemental data elements are also automatically included in the value set spreadsheet exported by the authoring tool.

CMS is evaluating additional sets for inclusion in the future, and these may include preferred language and socioeconomic status, among others.

Author narrative

It is important for measure developers to write a corresponding brief narrative description of a measure's supplemental data elements. The narrative descriptions for supplemental data elements about gender, race, ethnicity, and payer are automatically added to the metadata by the measure authoring tool.

Step 8: Define transmission format

In this step, the measure author describes how single patient or aggregate patient data will be communicated. Within the eMeasure, the measure author may insert URLs or hyperlinks that link to the transmission formats that are specified for a particular reporting program. For example, it could be a URL or a hyperlink that points to the Quality Reporting Document Architecture (QRDA) Category I

[10]The U.S. Centers for Disease Control and Prevention (CDC) has prepared a code set for use in coding race and ethnicity data. This code set is based on current federal standards for classifying data on race and ethnicity, specifically the minimum race and ethnicity categories defined by the U.S. Office of Management and Budget (OMB) and a more detailed set of race and ethnicity categories maintained by the U.S. Bureau of the Census (BC).
[11]Ibid.

implementation guide, or a hyperlink that points to the PQRI Registry XML specification. The measure author will not author the actual transmission format in the Measure Authoring Tool. What follows is an overview of different Transmission Formats.

There are several different ways of transmitting quality data. To transmit single patient-level quality data, developers can use the HL7 QRDA Category I as well as the HL7 Continuity of Care Document (CCD) standard. For aggregate-level patient quality data reporting, one can use HL7 QRDA Category II or Category III[12], or Physician Quality Reporting Initiative (PQRI) XML format[13]. The measure developer should consult with their COR/GTL to determine what transmission format should be used. For example, the program may have determined that all measures will use corresponding QRDA Category I reports to transmit patient data back to CMS.

The following are examples of transmission formats:

Single patient-level transmission format

These formats support the transmission of individual patient-level data.

CCD/C32

The purpose of CCD and Healthcare Information Technology Standards Panel (HISTP) C32 specifications is to provide a summary or snapshot of the status of a patient's health and healthcare in HL7 Clinical Document Architecture (CDA) format.

QRDA Category I

Though a CCD carries single patient data, it was designed to carry a specific set of summary data in support of a transition of care scenario. As a result, CCD will often contain some, but not all, of the data needed to determine whether or not a particular patient meets the population criteria within a particular measure. On the other hand, QRDA Category I was specifically designed to carry quality data tailored to a specific measure or measure set. As such, QRDA and CCD overlap considerably in data content, but not entirely. http://www.hl7.org/permalink/?CDAR2_QRDA

QRDA standardizes the representation of measure-defined data elements to enable interoperability between all of the stakeholder organizations. QRDA specifies a framework for quality reporting. A QRDA Category I report is an individual patient-level quality report; it contains raw applicable patient data for one patient and for one or more quality measures. A sample QRDA Category I report is shown in Figure 9-5.

QRDA reuses CCD templates wherever possible for its payload. The reuse of CCD templates in QRDA enables rapid implementation and development of QRDA and aligns with the eMeasure specification. Since all of the quality patterns in the Measure Authoring Tool pattern library are developed based on the HL7 V3 Reference Information Model (RIM) and the CCD specification is also developed from the

[12]QRDA Category II and QRDA Category III are draft specifications that have not been formally balloted at any level in HL7, and are therefore not described further in the Blueprint at this time.
[13]http://www.cms.gov/PQRS//Downloads/PQRI_2010RegistryXMLSpecifications.pdf

RIM, the quality patterns can be automatically converted to a corresponding CDA template for use in CCD (if within CCD's scope) and/or within QRDA.

Figure 9-5 QRDA Category I Sample Report

QRDA Category I Report

Patient	Ned Nuclear		
Date of birth	February 1, 1980	**Sex**	Male
Race	White	**Ethnicity**	Not Hispanic or Latino
Contact info	address not available Telecom information not available	**Patient IDs**	987654321 2.16.840.1.113883.19.5
Document Id	f2d5f971-d67a-4456-8833-213f01331ca0		
Document Created:	January 10, 2012		
Author	Quality Manager, RN, Good Health Clinic		

Table of Contents

- Measure Section
- Reporting Parameters
- Patient Data

Measure Section

- Measure: Surgery Patients with Appropriate Hair Removal (NQF0301)

Reporting Parameters

- Reporting period: 01 Jan 2012 - 31 Dec 2012

Patient Data

- Device, Applied: Hair Clipper
- Device, Applied: Razor
- Encounter: Encounter Inpatient
- Intervention, Performed: Hair Removal
- Intervention, Performed: Self Hair Removal
- Patient Characteristics: Clinical Trial Participant
- Patient Characteristics: Payer
- Procedure, Performed: SCIP Major Surgical Procedure

Document maintained by	Good Health Clinic
Informant	Good Health Clinic
Legal authenticator	Quality Manager, RN of Good Health Clinic signed at January 10, 2012

Aggregate-level transmission format

These formats support the transmission of data reflecting aggregate data or a summary of data from multiple patients.

PQRI XML Registry Specification

The Physician Quality Reporting Initiative (PQRI) XML Registry specification[14] was designed for aggregate reporting from standalone registry products. This specification was adopted by Office of the National Coordinator for Health Information Technology as the Stage 1 Meaningful Use standard for aggregate-level quality reporting. It is a government-unique standard, not a voluntary consensus standard.

QRDA Category III

Whereas a QRDA Category I document carries single patient data, a QRDA Category III document carries aggregate data, for one or more quality measures. At the time of this writing, the QRDA Category III specification is going through ballot in HL7.

[14] http://www.cms.gov/PQRS//Downloads/PQRI_2010RegistryXMLSpecifications.pdf

Step 9: Complete measure testing, document the measure, and obtain COR/GTL approval

Development of the complete technical specifications is an iterative process including input from the measure steward (for retooled measures), information gathering, TEP, public comments and measure testing. (Refer to the Measure Testing section for guidance on alpha and beta testing methods during measure development). Complete the detailed technical specifications and any additional documents required to produce the measure as it is intended (e.g., risk adjustment methodologies, etc.). The complete eMeasure specifications are documented in the NQF-developed MAT (for both retooled and de novo measures) and Measure Justification form (for de novo measures only). Information from measure testing, the public comment period, or other stakeholder input may result in the need to make changes to the technical specifications. The measure contractor will work with the TEP to incorporate these changes before submitting the final measure to the COR/GTL for approval.

The Measure Authoring Tool's output and Measure Justification form are designed as such to meet measure contractor's needs should the completed eMeasure be submitted to NQF for endorsement. The Measure Justification form, which is aligned with the NQF Measure Submission Form, was designed to guide the measure contractor throughout the measure development process in gathering the information in a standardized manner. The form also provides a crosswalk to the fields in the NQF Measure Submission Form to facilitate online information entry if CMS decides to submit the measure for consideration.

Step 10: Submit the eMeasure to NQF for endorsement

Refer to the National Quality Forum Endorsement section for further discussion of this process. When seeking NQF endorsement for an eMeasure, it is important to note that the preferred submission format is based on the Measure Authoring Tool output. Contractors should consult with the NQF Web site for any current policy decisions related to the endorsement of eMeasures.

NQF endorses measures only as a part of a larger project to seek consensus standards (measures) for a given health care condition or practice. If CMS decides that the measures developed under a project are to be submitted to the National Quality Forum (NQF) for endorsement, and NQF is conducting a project for which the measure is applicable, the measure contractor will assist CMS in the submission process. Measures are submitted to NQF using the Web-based electronic submission form. Upon the direction of the COR/GTL, the contractor will initiate the submission and complete the submission form.

NQF makes periodic updates to the measure submission process. Consult the NQF Web site below for the current process.

http://www.qualityforum.org/Measuring_Performance/Consensus_Development_Process.aspx

If the measure is submitted to NQF for endorsement, the measure contractor is required to provide technical support throughout the review process. This may include presenting the measure to the

Steering Committee that is evaluating the measure, answering questions from NQF staff or the Steering Committee about the specifications, testing or evidence.

During the course of the review, NQF may recommend revisions to the measure. If so, all subsequent revisions must be reported to and approved by the COR/GTL. Update the eMeasure in the Measure Authoring Tool with any changes agreed upon during the endorsement process.

The measure contractor for any measure developed under contract with CMS should list CMS as the steward, unless special arrangements have been made in advance. Developers should consult with the COR/GTL if there are questions on this. Barring special arrangements, the following format should be used—Center for Medicare and Medicaid Services (CMS).

9.8 Preparing eMeasure Specifications for Public Release

After the measure contractor completes eMeasures following the steps described above, the final eMeasure specifications will be submitted to CMS for public release. CMS posts the

1. Final measure specifications
2. Release notes
3. Logic guidance document
4. Guide to understanding and interpreting eMeasures
5. Calculation algorithms for eligible professional (EP) measures[15]
6. Resource table for EP measures[16]

at http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/ClinicalQualityMeasures.html.

eMeasure Naming Conventions

CMS measure contractors must follow the conventions detailed below when submitting final measure specifications for public release on the CMS Web site.

CMS eMeasure Identifier

Beginning with the release of 2014 clinical quality measure specifications, CMS created a unique "CMS eMeasure Identifier" to clearly and consistently identify eMeasure files. It is a concatenation of the eMeasure identifier assigned to the measure in the MAT with the "eMeasure Version Number", which is prepended by "CMS". Based on this universal naming convention, the MU2 EP measure NQF0034 would display the following: CMS130v1.

eMeasure Information	Value
eMeasures Identifier (MAT)	130

[15]Also available: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Electronic_Reporting_Spec.html.
[16]Ibid.

eMeasure Information	Value
eMeasure Version Number	1
CMS eMeasure ID	**CMS130v1**

Individual eMeasure Files

The file type (.xml or .html) is added to the CMS eMeasure ID to complete the naming convention for the components of the eMeasure package (example: CMS130v1.xml).

Packaging File Names

The naming conventions for the individual eMeasure packages (zip files and measure folder) that will contain the eMeasure XML file and human-readable rendition are described below in the order in which they must appear.

1. Setting for which the measure applies—"EP" or "EH" measures
2. CMS eMeasure ID
3. NQF identifier—if not endorsed by NQF, the file will contain "NQFXXXX"
4. Abbreviated name for the clinical quality measure (example: "Colorectal_Cancer_Screen")

Using the EP measure example from above for NQF0034, the Zip file and corresponding folder would appear as follows:

Artifact	File Name
Measure Zip file	EP_CMS130v1_NQF0034_Colorectal_Cancer_Screen.zip
Measure folder	EP_CMS123v1_NQF0056_ Colorectal_Cancer_Screen

eSpecification "eSpec" Navigator

The eSpecification "eSpec" Navigator is a web-based application that allows easy browsing of the eMeasure specifications files or artifacts[17]. It provides access to both EP and eligible hospital (EH) eMeasures for 2014. A user can download eMeasure packages, search, browse, and sort their contents using the eSpec Navigator. The eSpec Navigator also provides direct links to NLM's Value Set Authority Center for each value set that is used by the eMeasures. QRDA Category I sample files for each eMeasure are also available on the eSpec Navigator. Refer to the Help page of the eSpec Navigator for instructions about the tool.

Guide for Reading eMeasures

CMS published the *Guide for Reading Eligible Professional (EP) and Eligible Hospital (EH) eMeasures* in October 2012 to provide guidance in understanding and implementing the EP and EH eMeasures

[17]eSpecification "eSpec" Navigator. http://www.lantanagroup.com/especnavigator/#home.

specified for 2014. The guide contains useful information for understanding the human readable rendition (HTML) file of an eMeasure. The guide is available for download from the CMS web site[18].

eMeasure Versioning

Once CMS has released an eMeasure for public use, if any changes are made to the measure's specifications (e.g., major, minor, technical corrections, annual update), the version numbers will be advanced to the next whole number (e.g., from 1 to 2) prior to posting.

9.9 eMeasure Future and Next Steps

The HQMF standard is currently an HL7 Draft Standard for Trial Use, meaning that it was balloted as a draft so that it could be tested for finalization. Following a one- to two-year testing period, HQMF will be taken back through the HL7 process to turn it into an official ANSI-accredited standard. Measure developers are encouraged to post comments about the HQMF standard through this link:

http://www.hl7.org/dstucomments/showdetail.cfm?dstuid=39. The comments will be reviewed when HQMF goes back through ballot.

A U.S. Realm HL7 Implementation Guide for eMeasure is currently under development and will be balloted through HL7. This Implementation Guide will detail how to develop eMeasures based on the QDM and the building-block approach.

QRDA Category I is also an HL7 DSTU. QRDA Category III is a draft specification currently going through ballot in HL7.

[18]Guide for Reading Eligible Professional (EP) and Eligible Hospital (EH) eMeasures. http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/Guide_Reading_EP_Hospital_eCQMs.pdf

Appendix 9a[19] XML View of a Sample eMeasure[20]

This appendix shows a detailed example of an eMeasure, encoded per the HQMF standard and the additional rules stipulated in this section, illustrating how it will appear when rendered in a web browser. Following the example, a high-level view of the underlying XML is provided.

eMeasure Title	Asthma Assessment		
eMeasure Identifier (Measure Authoring Tool)	123	**eMeasure Version Number**	1
NQF Number	0001	**GUID**	59B24505-B9D1-48B8-BEB2-AF35219AA689
Measurement Period	January 1, 2011 through December 31, 2011		
Measure Steward	American Medical Association-Physician Consortium for Performance Improvement		
Measure Developer	American Medical Association-Physician Consortium for Performance Improvement		
Endorsed by	National Quality Forum		
Description	Percentage of patients aged 5 through 40 years with a diagnosis of asthma who were evaluated during at least one office visit within 12 months for the frequency (numeric) of daytime and nocturnal asthma symptoms.		
Copyright	©2010 American Medical Association. All Rights Reserved		
Disclaimer	None		

[19]Note that though value sets are authored in the measure authoring tool, they are not part of the published eMeasure. Rather, the published eMeasure contains value set references, and the value sets themselves are exported from the tool in a spreadsheet.
[20]This is a sample measure, solely for illustration purposes. It is not a real measure.

Measure scoring	Proportion
Measure type	Process
Stratification	Population 1: DOB 14-23 years before start of measurement period Population 2: DOB 14-19 years before start of measurement period Population 3: DOB 20-23 years before start of measurement period
Risk Adjustment	None
Rate Aggregation	None
Rationale	Appropriate treatment of asthma patients requires accurate classification of asthma severity. Physician assessment of the frequency of asthma symptoms is the first step in classifying asthma severity.
Clinical Recommendation Statement	To determine whether the goals of therapy are being met, monitoring is recommended in the 6 areas listed below: ◆ Signs and symptoms (daytime; nocturnal awakening) of asthma ◆ Pulmonary function (spirometry; peak flow monitoring) ◆ Quality of life/functional status ◆ History of asthma exacerbations ◆ Pharmacotherapy (as-needed use of inhaled short-acting beta2-agonist, adherence to regimen of long-term-control medications) ◆ Patient-provider communication and patient satisfaction (NAEPP/NHLBI)
Improvement Notation	Higher score indicates better quality
Reference	National Asthma Education and Prevention Program Expert Panel Report 2: Guidelines for the diagnosis and management of asthma. National Heart, Lung, and Blood Institute, National Institutes of Health; July 1997. NIH Publication No. 97-4051. Available at: http://www.nhlbi.nih.gov/guidelines/asthma/asthgdln.htm. Accessed August 2002.

Reference	National Asthma Education and Prevention Program Expert Panel Report 2 Update: Guidelines for the diagnosis and management of asthma – update on selected topics 2002. National Heart, Lung, and Blood Institute, National Institutes of Health; August 2002. NIH Publication No. 97-4051. Available at: http://www.nhlbi.nih.gov/guidelines/asthma/asthsumm.htm. Accessed September 2002.
Definition	None
Guidance	None
Transmission Format	None
Initial Patient Population	(Brief narrative description of population. Must be consistent with the computer-generated narrative below. The computer generated narrative is standardized and concise, and can lack the richness of full text that sometimes helps in the understanding of a measure. This is especially true for measures that have complex criteria, where the computer generated text may not be able to express the exact description that a measure developer would like to convey. As part of the quality assurance step, it is important to compare the manually authored narrative against the automatically rendered narrative for any discrepancies).
Denominator	(Brief narrative description of population. If population isn't applicable, then say "Not Applicable". If population isn't defined, then say "None").
Denominator Exclusions	(Brief narrative description of population. If population isn't applicable, then say "Not Applicable". If population isn't defined, then say "None").
Numerator	(Brief narrative description of population. If population isn't applicable, then say "Not Applicable". If population isn't defined, then say "None").
Numerator Exclusions	Not Applicable
Denominator Exceptions	(Brief narrative description of population. If population isn't applicable, then say "Not Applicable". If population isn't defined, then say "None").
Measure Population	Not Applicable

Measure Observations	(Brief narrative description of measure observations. If not a Continuous Variable measure, then should say "Not Applicable").
Supplemental Data Elements	For every patient evaluated by this measure also identify payer, race, ethnicity and sex.

Table of Contents

- Population Criteria
- Data Criteria
- Measure Observations (only shown if present)
- Reporting Stratification (only shown if present)
- Supplemental Data Elements

Population criteria[21]

- Initial Patient Population =
  - AND:
    - AND: "Patient Characteristic Birthdate: birth date" >= 5 year(s) starts before start of "Measurement Period"
    - AND: "Patient Characteristic Birthdate: birth date" <= 40 year(s) starts before start of "Measurement Period"
  - AND: "Diagnosis Active: Asthma" starts before or during ("Encounter, Performed: Encounter Office & Outpatient Consult" during "Measurement Period")
  - AND: >= 2 count(s) of
    - AND: "Encounter, Performed: Encounter Office & Outpatient Consult" during "Measurement Period"
- Denominator =
  - AND: "Initial Patient Population"
- Denominator Exclusions =
  - None
- Numerator =
  - AND:

[21]Populations will vary based on type of measure scoring.

- OR:
  - AND: "Symptom Assessed: Asthma Daytime Symptoms Quantified"
  - AND: "Symptom Assessed: Asthma Nighttime Symptoms Quantified"
  - starts before or during ("Encounter, Performed: Encounter Office & Outpatient Consult" during "Measurement Period")
- OR:
  - AND: "Symptom Active: Asthma Daytime Symptoms"
  - AND: "Symptom Active: Asthma Nighttime Symptoms"
  - starts before or during ("Encounter, Performed: Encounter Office & Outpatient Consult" during "Measurement Period")
- OR: "Risk category / assessment: Asthma Symptom Assessment Tool" starts before or during ("Encounter: Encounter Office & Outpatient Consult" during "Measurement period")

- **Denominator Exceptions =**
  - None

Data criteria

- "Diagnosis, Active: Asthma" using "Asthma GROUPING Value Set (2.16.840.1.113883.3.526.03.362)"
- "Encounter, Performed: Encounter Office & Outpatient Consult" using "Encounter Office & Outpatient Consult CPT Value Set (2.16.840.1.113883.3.526.02.99)"
- "Patient Characteristic Birthdate: birth date" (age) using "birth date LOINC Value Set (2.16.840.1.113883.3.560.100.4)"
- "Risk Category / assessment: Asthma Symptom Assessment Tool" using "Asthma Symptom Assessment Tool SNOMED-CT Value Set (2.16.840.1.113883.3.526.2.143)"
- "Symptom, Active: Asthma Daytime Symptoms" using "Asthma Daytime Symptoms SNOMED-CT Value Set (2.16.840.1.113883.3.526.02.440)"
- "Symptom, Active: Asthma Nighttime Symptoms" using "Asthma Nighttime Symptoms SNOMED-CT Value Set (2.16.840.1.113883.3.526.2.441)"
- "Symptom, Assessed: Asthma Daytime Symptoms Quantified" using "Asthma Daytime Symptoms Quantified SNOMED-CT Value Set (2.16.840.1.113883.3.526.2.141)"
- "Symptom, Assessed: Asthma Nighttime Symptoms Quantified" using "Asthma Nighttime Symptoms Quantified SNOMED-CT Value Set (2.16.840.1.113883.3.526.2.142)"

Reporting stratification

- Reporting Stratum 1 =
  - AND: "Patient Characteristic Birthdate: birth date" >= 14 year(s) starts before start of "Measurement Period"
  - AND: "Patient Characteristic Birthdate: birth date" <= 23 year(s) starts before start of "Measurement Period"

- Reporting Stratum 2 =
  - AND: "Patient Characteristic Birthdate: birth date" >= 14 year(s) starts before start of "Measurement Period"
  - AND: "Patient Characteristic Birthdate: birth date" <= 19 year(s) starts before start of "Measurement Period"
- **Reporting Stratum 3 =**
  - AND: "Patient Characteristic Birthdate: birth date" >= 20 year(s) starts before start of "Measurement Period"
  - AND: "Patient Characteristic Birthdate: birth date" <= 23 year(s) starts before start of "Measurement Period"

Supplemental data elements

- "Patient Characteristic Ethnicity: Ethnicity" using "Ethnicity CDC Value Set (2.16.840.1.114222.4.11.837)"
- "Patient Characteristic Sex: Sex" using "ONC Administrative Sex (2.16.840.1.113762.1.4.1)"
- "Patient Characteristic Payer: Payer" using "Payer Source of Payment Typology Value Set (2.16.840.1.114222.4.11.3591)"
- "Patient Characteristic Race: Race" using "Race CDC Value Set (2.16.840.1.114222.4.11.836)"

Measure set CLINICAL QUALITY MEASURE SET 2011-2012

Appendix 9b Electronic Specification (eMeasure)

The prior appendix shows a detailed example of an eMeasure, encoded per the HQMF standard and the additional rules stipulated in this section, illustrating how it will appear when rendered in a web browser. This appendix provides a high-level view of the underlying XML.

The NQF-developed Measure Authoring Tool will generate a much more detailed XML document than is shown here, but measure developers will require some knowledge of XML and the HQMF standard in order to make subsequent manual edits to eMeasures generated by the tool.

Skeletal Example That Shows Major Components of a Prototypic eMeasure Document

```
<QualityMeasureDocument>
 ... eMeasure Header ...
 <section>
 <title>Data criteria</title>
 <text>... narrative data criteria descriptions ...</text>
 <entry>... Formal data criteria definition ...</entry>
 <entry>... Formal data criteria definition ...</entry>
 ...
 </section>
 <section>
 <title>Population criteria</title>
 <text>... narrative population criteria descriptions ...</text>
 <entry>... Formal population criteria definition ...</entry>
 <entry>... Formal population criteria definition ...</entry>
 ....
 </section>
 <section>
 <title>Measure observations</title>
 <text>... narrative measure observation descriptions ...</text>
 <entry>... Formal measure observation definition ...</entry>
 <entry>... Formal measure observation definition ...</entry>
 ...
```

```
</section>
<section>
<title>Reporting stratification</title>
<text>... narrative reporting stratification descriptions ...</text>
<entry>... Formal reporting stratification definition ...</entry>
<entry>... Formal reporting stratification definition ...</entry>
...
</section>
<section>
<title>Supplemental data elements</title>
<text>... narrative supplemental data elements descriptions ...</text>
<entry>... Formal supplemental data elements definition ...</entry>
<entry>... Formal supplemental data elements definition ...</entry>
...
</section>
<section>
<section>...</section>
</section>
</QualityMeasureDocument>
```

Appendix 9c1 Time Unit and Time Interval Definitions

ISO 8601:2004 defines data elements and interchange formats for the representation of dates and times, including time intervals. Below is a summary of important terms defined in the standard that are of particular importance and can be drawn upon to be used in time interval calculations for eMeasures.

Term	ISO Definition	ISO Notes
Time interval	Part of the time axis limited by two instants.	• A time interval comprises all instants between the two limiting instants and, unless otherwise stated, the limiting instants themselves.
Duration	Non-negative quantity attributed to a time interval, the value of which is equal to the difference between the time points of the final instant and the initial instant of the time interval, when the time points are quantitative marks.	• In case of discontinuities in the time scale, such as a leap second or the change from winter time to summer time and back, the computation of the duration requires the subtraction or addition of the change of duration of the discontinuity. • The SI unit of duration is the second.
Nominal duration	Duration expressed amongst others in years, months, weeks or days.	• The duration of a calendar year, a calendar month, a calendar week or a calendar day depends on its position in the calendar. Therefore, the exact duration of a nominal duration can only be evaluated if the duration of the calendar years, calendar months, calendar weeks or calendar days used are known.
Second	Base unit of measurement of time in the International System of Units (SI) as defined by the International Committee of Weights and Measures	
Minute	Unit of time, equal to 60 seconds	

Term	ISO Definition	ISO Notes
Hour	Unit of time, equal to 60 minutes	
Day <unit of time>	Unit of time, equal to 24 hours	
Calendar day	Time interval starting at midnight and ending at the next midnight, the latter being also the starting instant of the next calendar day	• A calendar day is often also referred to as a day. • **The duration of a calendar day is 24 hours**; except if modified by: ○ the insertion or deletion of leap seconds, by decision of the International Earth Rotation Service (IERS) or ○ the insertion or deletion of other time intervals, as may be prescribed by local authorities to alter the time scale of local time.
Day <duration>	Duration of a calendar day.	• The term "day" applies also to the duration of any time interval which starts at a certain time of day at a certain calendar day and ends at the same time of day at the next calendar day.
Calendar week	Time interval of seven calendar days starting with a Monday.	• A calendar week is also referred to as week.
Week	Duration of a calendar week.	• The term "week" also applies to the duration of any time interval which starts at a certain time of day at a certain calendar day and ends at the same time of day at the same calendar day if the next calendar week.
Calendar month	Time interval resulting from the division of a calendar year in 12 time intervals, each with a specific name and containing a specific number of calendar days.	• A calendar month is often referred to as month.

Term	ISO Definition	ISO Notes
Month	Duration of 28, 29, 30 or 31 calendar days depending on the start and/or the end of the corresponding time interval within the specific calendar month.	• The term "month" applies also to the duration of any time interval which starts at a certain time of day at a certain calendar day of the calendar month and ends at the same time of day at the same calendar day of the next calendar month, if it exists. In other cases the ending calendar day has to be agreed on. • In certain applications a month is considered as a duration of 30 calendar days.
Calendar year	Cyclic time interval in a calendar which is required for one revolution of the Earth around the Sun and approximated to an integral number of calendar days.	• A calendar year is also referred to as year. • Unless otherwise specified the term designates in this International Standard a calendar year in the Gregorian calendar.
Year	Duration of 365 or 366 calendar days depending on the start and/or the end of the corresponding time interval within the specific calendar year.	• The term "year" applies also to the duration of any time interval which starts at a certain time of day at a certain calendar date of the calendar year and ends at the same time of day at the same calendar date of the next calendar year, if it exists. In other cases the ending calendar day has to be agreed on.
Common year	Calendar year in the Gregorian calendar that has 365 calendar days.	
Leap year	Calendar year in the Gregorian calendar that has 366 days.	

The standard also defines multiple formats for the representation of date and time as well as time intervals and durations. ISO 8601 postulates that duration can be expressed by a combination of components with accurate duration (hour, minute and second) and components with nominal duration (year, month, week and day). The standard allows for the omission of lower level components for "reduced accuracy" applications.

Appendix 9c2 Time Interval Calculation Conventions

The unit in which a time interval (or its duration) is expressed may be dependent upon the necessary level of accuracy for the purposes of measurement. As such, the selection of the time unit to be used should be made according to the level of granularity needed to meet the intent of the measure.

These conventions are intended to standardize the interpretation of time calculation units for durations, e.g., difference between two date/time elements, typically with time relationships defined in the QDM, such as "starts after start of" and "ends before start of".

Calculation Unit	Definition	Calculation[22,23,24]
Year	Duration of any time interval which starts at a certain time of day at a certain calendar date of the calendar year and ends at: • the same time of day at the same calendar date of the next calendar year, if it exists; • the same time of day at the immediately following calendar date of the ending calendar year, if the same calendar date of the ending calendar year does not exist **Notes:** 1. When in the next calendar year the same calendar date does not exist, the ISO states that the ending calendar day has to be agreed upon. The above convention is proposed as a resolution to this issue. 2. The ISO permits the representation of dates and times with "reduced accuracy", when necessary. For the purposes of quality measures, duration expressed in years ignores the time of day associated with the date/time stamps used in the calculation, e.g., the number of years will not change until the month and day of	1. **Month (date 2) < month (date 1):** Duration (years) = year (date 2) – year (date 1) – 1 <u>Example 1:</u> Date 1: 2012-**03**-10 22:05:09 Date 2: 2013-**02**-18 19:10:03 Duration = year (date 2) – year (date 1) – 1 = 2013 – 2012 – 1 **= 0 years** 2. **Month (date 2) = month (date 1) <u>and</u> day (date 2) >= day (date 1)** Duration (years) = year (date 2) – year (date 1) <u>Example 2.a:</u> day (date 1) = day (date 2) Date 1: 2012-03-**10** 22:05:09 Date 2: 2013-03-**10** 08:01:59

[22]For the purposes of this document, date 2 is assumed to be more recent than date 1.
[23]All dates are represented according to the extended formatted designated in ISO 8601: YYYY-MM-DD HH:MM:SS and a 24 hour timekeeping system.
[24]The result of the calculation is always an integer; if the result includes decimals, it should be truncated (rounded down) to the unit.

Calculation Unit	Definition	Calculation[22,23,24]
	date 2 reach (or surpass) the month and day of date 1; time of day should be ignored or set to 00:00:00 for this calculation.	Duration = year (date 2) – year (date 1) = 2013 – 2012 **= 1 year** **Note:** Time of day is ignored in this calculation. If time of day were to be considered for the calculation, the duration of the time interval would be 0 years according to the definition. <u>Example 2.b:</u> day (date 2) > day (date 1) Date 1: 2012-03-**10** 22:05:09 Date 2: 2013-03-**20** 04:01:30 Duration = year (date 2) – year (date 1) = 2013 – 2012 **= 1 year** **3. Month (date 2) = month (date 1) <u>and</u> day (date 2) < day (date 1)** Duration (years) = year (date 2) – year (date 1) – 1 <u>Example 3.a:</u> Date 1: 2012-02-**29** Date 2: 2014-02-**28** Duration = year (date 2) – year (date 1) – 1 = 2014 – 2012 – 1 = **1 year** **4. Month (date 2) > month (date 1)** Duration (years) = year (date 2) – year (date 1) <u>Example 4.a:</u> Date 1: 2012-**03**-10 11:16:02 Date 2: 2013-**08**-15 21:34:16 Duration = year (date 2) – year (date 1) = 2013 – 2012 **= 1 year**

Calculation Unit	Definition	Calculation[22,23,24]
		Example 4.b: Date 1: 2012-**02**-29 10:18:56 Date 2: 2014-**03**-01 19:02:34 Duration = year (date 2) – year (date 1) = 2014 – 2012 **= 2 years** **Note:** Because there is no February 29 in 2014, the number of years can only change when the date reaches March 1st, the first date in 2014 that surpasses the month and day of date 1 (February 29).
Month	Duration of any time interval which starts at a certain time of day at a certain calendar day of the calendar month and ends at: • the same time of day at the same calendar day of the ending calendar month, if it exists; • the same time of day at the immediately following calendar date of the ending calendar month, if the same calendar date of the ending month year does not exist **Notes:** 1. When in the next calendar year the same calendar date does not exist, the ISO states that the ending calendar day has to be agreed upon. The above convention is proposed as a resolution to this issue. 2. The ISO permits the representation of dates and times with "reduced accuracy", when necessary. For the purposes of quality measures, duration expressed in months ignores the time of day associated with the date/time stamps used in the calculation, e.g., the number of months will not change until the month and day of date 2 reach (or surpass) the month and day of date 1; time of day should be ignored or set to 00:00:00 for this calculation.	**1. Day (date 2) >= day (date 1)** Duration (months) = (year (date 2) – year (date 1))*12 + (month (date 2) – (month date 1)) Example 1.a: Date 1: 2012-03-01 14:05:45 Date 2: 2012-03-31 23:01:49 Duration = (year (date 2) – year (date 1))*12 + (month (date 2) – (month date 1)) = (2012 – 2012)*12 + (3 – 3) **= 0 months** Example 1.b: Date 1: 2012-03-10 22:05:09 Date 2: 2013-06-30 13:00:23 Duration = (year (date 2) – year (date 1))*12 + (month (date 2) – (month date 1)) = (2013 – 2012)*12 + (6 – 3) = 12 + 3 **= 15 months**

Calculation Unit	Definition	Calculation[22,23,24]
		2. Day (date 2) < day (date 1) Duration (months) = (year (date 2) – year (date 1))*12 + (month (date 2) – month (date 1)) – 1 <u>Example 2:</u> Date 1: 2012-03-10 22:05:09 Date 2: 2013-01-29 07:19:33 Duration = (year (date 2) – year (date 1))*12 + (month (date 2) – (month date 1)) – 1 = (2013 – 2012)*12 + (1 – 3) = 12 – 2 **= 10 months**
Week	Duration of any time interval which starts at a certain time of day at a certain calendar day at a certain calendar week and ends at the same time of day at the same calendar day of the ending calendar week. **Notes:** 1. The ISO permits the representation of dates and times with "reduced accuracy", when necessary. For the purposes of quality measures, duration expressed in months ignores the time of day associated with the date/time stamps used in the calculation, e.g., a complete week is always seven days long: a week has passed when the same weekday in the following week is reached; time of day should be ignored or set to 00:00:00 for this calculation.	**1. Duration = [date 2 – date 1 (days[25])]/7** <u>Example 1:</u> Date 1: 2012-03-10 22:05:09 Date 2: 2012-03-20 07:19:33 Duration = [# days (month (date 1)) – day (date 1) + # days (month (date 1) + 1) + # days (month (date 1) + 2) + … + # days (month (date 2) – 1) + day (date 2)]/7 = (20 – 10)/7 = 10/7 **= 1 week** (result truncated to integer).

[25]For information on how to calculate the duration, in days, of a time interval, please see "day" below.

Calculation Unit	Definition	Calculation[22,23,24]
Day	Duration of any time interval which starts at a certain calendar day and ends at the next calendar day (1 second to 47 hours, 59 minutes and 59 seconds); the number of times midnight is crossed. **Notes:** 1. The ISO permits the representation of dates and times with "reduced accuracy", when necessary. For the purposes of quality measures, duration expressed in months ignores the time of day associated with the date/time stamps used in the calculation, e.g., the number of years will not change until the month and day of date 2 reach (or surpass) the month and day of date 1; time of day should be ignored or set to 00:00:00 for this calculation.	The duration in days between two dates, will be generally given by subtracting the start calendar date to the end calendar date, regardless of the time of day between the two dates. <u>Example 1:</u> Date 1: 2012-01-31 12:30:00 Date 2: 2012-02-01 09:00:00 Duration = 02-01 – 01-31 **= 1 day** <u>Example 2:</u> Date 1: 2012-01-31 12:30:00 Date 2: 2012-02-01 14:00:00 Duration = 02-01 – 01-31 **= 1 day**
Hours	The number of 60 minute cycles between two given dates. **Notes:** 1. Use the date, hour and minute of the date/time stamps to compute the interval. 2. Seconds are not used in the calculation.	The duration in hours between two dates is the number of minutes between the two dates, divided by sixty. The result is truncated to the unit. <u>Example 1:</u> Date/Time 1: 2012-03-01 03:10 Date/Time 2: 2012-03-01 05:09 Duration **= 1 hour** <u>Example 2:</u> Date/Time 1: 2012-02-29 23:10 Date/Time 2: 2012-03-01 00:10 Duration **= 1 hour** <u>Example 3:</u> Date/Time 1: 2012-03-01 03:10 Date/Time 2: 2012-03-01 04:00 Duration = **0 hours**

Calculation Unit	Definition	Calculation[22,23,24]
Minutes	The number of minutes between two given dates. **Notes:** 1. Seconds are **not** used in the calculation.	Example 1: Date/Time 1: 2012-03-01 03:10 Date/Time 2: 2012-03-01 05:20 Duration **= 130 minutes** Example 2: Date/Time 1: 2012-02-29 23:10 Date/Time 2: 2012-03-01 00:20 Duration **= 70 minutes**

Timing relationships with no calculation unit defined

When no threshold is defined in a measure phrase that compares timing between two elements, a common level of granularity needs to be established for comparisons within and across EHRs. For instance, the criterion:

A starts before start of B

Depending on the level of granularity of the data (e.g. HH:MM:SS vs. HH:MM), the result of the timing comparison may vary if no comparison unit is defined:

Example 1

A: 2012-01-01 11:00:01

B: 2012-01-01 11:00:02

A starts before B = TRUE

Example 2

A: 2012-01-01 11:00

Appendix 9d Defintions

General Definitions	
eMeasure	An eMeasure is a health quality measure encoded in a health quality measure format (HQMF). See HQMF.
Health Quality Measures Format (HQMF)	A standards-based representation of quality measures. A quality measure expressed in HQMF format is also referred to as an "eMeasure".
Quality Measure	A quality measure is in effect a rule (or the result of a rule) that assigns numeric values to a specific quality indicator, usually in relation to a specified process or outcome via the measurement of an action, process or outcome of clinical care. Quality measures generally consist of a descriptive statement or indicator, a list of data elements that are necessary to construct and/or report the measure, detailed specifications that direct how the data elements are to be collected (including the source of data), the population on whom the measure is constructed, the timing of data collection and reporting, and the methods used to construct the measure.
Quality Measure Set	A unique grouping of performance measures carefully selected to provide, when viewed together, a general picture of the care provided in a given domain (e.g., cardiovascular care, pregnancy).
Measure Parameter Definitions	
Refer to Table 1	
Quality Measure Scoring	
Continuous Variable	A measure score in which each individual value for the measure can fall anywhere along a continuous scale, and can be aggregated using a variety of methods such as the calculation of a mean or median (e.g., mean number of minutes between presentation of chest pain to the time of administration of thrombolytics).
Proportion	A score derived by dividing the number of cases that meet a criterion for quality (the numerator) by the number of eligible cases within a given time frame (the denominator) where the numerator cases are a subset of the denominator cases (e.g., percentage of eligible women with a mammogram performed in the last year).
Ratio	A score that may have a value of zero or greater that is derived by dividing a count of one type of data by a count of another type of data (e.g., the number of patients with central lines who develop infection divided by the number of central line days).

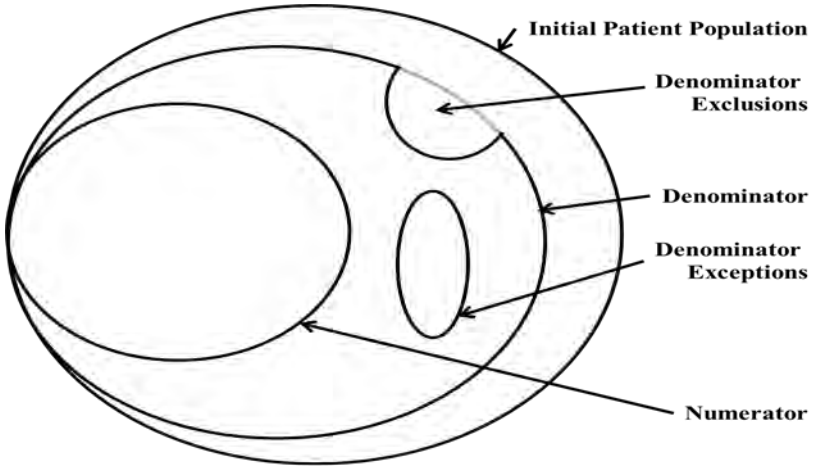
Quality Measure Types	
Outcome measure	A measure that indicates the result of the performance (or nonperformance) of functions or processes. A measure that focuses on achieving a particular state of health.
Process measure	A measure that focuses on a process which leads to a certain outcome, meaning that a scientific basis exists for believing that the process, when executed well, will increase the probability of achieving a desired outcome.

Appendix 9e Proportion eMeasure Calculations

This appendix provides further guidance on the precise mathematical relationships between populations in a proportion measure, and the process to be used to determine individual and aggregate scores. This ensures that all implementers come up with the same scores, given the same data and same eMeasures.

The figure below shows a fixed mathematical relationship between the populations in a proportion measure. The definitions of the various populations in a Proportion Measure are listed in Table 9-1 (located at the beginning of this section).

Proportion Measure Populations



From these relationships and definitions, we define the sequential steps to be used when determining whether or not a patient falls into a given population:

1. **Initial Patient Population**: Identify those patients that meet the IPP criteria.
2. **Denominator**: Identify that subset of the IPP that meet the DENOM criteria.
3. **Denominator Exclusions**: Identify that subset of the DENOM that meet the EXCL criteria.
4. **Numerator**: Identify those in the DENOM and NOT in the EXCL that meet the NUMER criteria.
5. **Denominator Exceptions**: Identify those in the DENOM and NOT in the EXCL and NOT in the NUMER that meet the EXCEP criteria.

Queries should be based on the principle of "**positive evidence**". Positive evidence is defined as data that can be used to confirm that a given criterion was met. The principle is particularly relevant where there is no data, or where there is conflicting data. Where, for instance, a NUMER criterion is "LDL Cholesterol is less than 100" and there is no LDL Cholesterol result in the EHR, then there is no positive evidence, and the criterion is not met. Where, for instance, a DENOM criterion is "ejection fraction is

less than 40", and there is both an ejection fraction of less than 40 and an ejection fraction of greater than 40 in the EHR, then because there is positive evidence of an ejection fraction less than 40, the criterion is met.[26]

Proportion measure example

A fictitious proportion measure defines the following population criteria:

IPP: all patients aged 65 years and older with an active diagnosis of diabetes mellitus.

DENOM: equals IPP.

EXCL: bilateral blindness

NUMER: dilated eye exam for diabetic retinopathy

EXCEP: bed confinement status in a community where mobile eye-exam imaging is unavailable

eMeasure individual determination:

Mr. Jones is 75 years old, with an active diagnosis of diabetes. There is no mention of blindness in his chart. He has a documented dilated eye exam for diabetic retinopathy.

(IPP = YES) Mr. Jones meets the IPP criteria.

(DENOM = YES) Mr. Jones meets the DENOM criteria.

(EXCL = NO) By the "positive evidence" principle, Mr. Jones does not meet the EXCL criteria.

(NUMER = YES) Mr. Jones meets the NUMER criteria.

(EXCEP = NO) By definition, Mr. Jones does not meet the EXCEP criteria, because EXCEP criteria are not applicable to those meeting the NUMER criteria.

Mr. Smith is 75 years old, with an active diagnosis of diabetes. There is no mention of blindness in his chart. There is no mention of dilated eye exam in his chart. There is no mention in his chart that he is bed bound.

(IPP = YES) Mr. Smith meets the IPP criteria.

(DENOM = YES) Mr. Smith meets the DENOM criteria.

(EXCL = NO) By the "positive evidence" principle, Mr. Smith does not meet the EXCL criteria.

(NUMER = NO) By the "positive evidence" principle, Mr. Smith does not meet the NUMER criteria.

(EXCEP = NO) By the "positive evidence" principle, Mr. Smith does not meet the EXCEP critieria.

[26]Many measures will be more specific with respect to which observation to use when comparing against a criterion (such as "**LAST** ejection fraction is less than 40").

Mr. Johnson is 85 years old, with an active diagnosis of diabetes. There is no mention of blindness in his chart. He has a documented dilated eye exam for diabetic retinopathy. He is known to be confined to bed in a community where mobile eye-exam imaging is unavailable.

(IPP = YES) Mr. Johnson meets the IPP criteria.

(DENOM = YES) Mr. Johnson meets the DENOM criteria.

(EXCL = NO) By the "positive evidence" principle, Mr. Johnson does not meet the EXCL criteria.

(NUMER = YES) Mr. Johnson meets the NUMER criteria.

(EXCEP = NO) By definition, Mr. Johnson does not meet the EXCEP criteria, because EXCEP criteria are not applicable to those meeting the NUMER criteria.

eMeasure aggregate calculations:

Aggregate scores are simply the counts of individuals in each population. Thus, the aggregate IPP is the count of individuals meeting the IPP criteria; the aggregate DENOM is the count of individuals meeting the DENOM criteria; the aggregate EXCL is the count of individuals meeting the EXCL criteria; the aggregate NUMER is the count of individuals meeting the NUMER criteria; the aggregate EXCEP is the count of individuals meeting the EXCEP criteria.

The "**performance rate**" is a ratio of patients meeting NUMER criteria, divided by patients in the DENOM (accounting for exclusions and exceptions). Performance Rate can be calculated using this formula:

Performance Rate = (NUMER) / (DENOM – EXCL – EXCEP)

From the example above, counting all individuals within the population, the following aggregate counts are determined:

Initial Patient Population: N=150 (e.g. 150 patients meet the IPP criteria).

Denominator: N=150.

Denominator Exclusions: N=20 (meet DENOM and also meet EXCL)

Numerator: N=75 (meet DENOM, not in EXCL, and also meet NUMER criteria)

Denominator Exceptions: N=5 (meet DENOM, not in EXCL, not in NUMER, and also meet the EXCEP criteria).

Performance Rate = (NUMER) / (DENOM – EXCL – EXCEP) = (75)/(150-20-5) = 0.6

Appendix 9f Acronyms and Abbreviations

ANSI	American National Standards Institute
ASTM	*ASTM* International, originally known as the *American Society for Testing and Materials*
CCD	Continuity of Care Document
CCR	Continuity of Care Record
CDA	Clinical Document Architecture
CMS	Centers for Medicare & Medicaid Services
CPT	Current Procedural Terminology
CQM	Clinical Quality Measures
CVX	Vaccines Administered
CPT-4	Current Procedural Terminology, Fourth Edition
DSTU	Draft Standard for Trial Use
EHR	Electronic Health Record
GUI	Graphical User Interface
HCPCS	Healthcare Common Procedure Coding System
HHS	Department of Health and Human Services
HITPC	Healthcare Information Technology Policy Committee
HITSC	Healthcare Information Technology Standards Committee
HITSP	Healthcare Information Technology Standards Panel
HL7	Health Level Seven

HQMF	Health Quality Measures Format
ICD-9-CM	International Classification of Diseases, Ninth Revision, Clinical Modification
ICD-10-CM	International Classification of Diseases, Tenth Revision, Clinical Modification
ICD-10-PCS	International Classification of Diseases, Tenth Revision, Procedural Coding System
IG	Implementation Guide
IPP	Initial Patient Population
LOINC	Logical Observation Identifiers Names and Codes
MAT	Measure Authoring Tool
NLM	National Library of Medicine
NQF	National Quality Forum
ONC	Office of the National Coordinator for Health Information Technology
PHDSC	Public Health Data Standards Consortium
PHIN-VADS	CDC Public Health Information Network (PHIN) Vocabulary Access and Distribution System (VADS)
PQRI	Physician Quality Reporting Initiative
PQRS	Physician Quality Reporting System
QDM	Quality Data Model (QDM, previously known as Quality Data Set)
QRDA	Quality Reporting Document Architecture
RIM	HL7 V3 Reference Information Model

SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
UCUM	Unified Code for Units of Measure
USCRS	SNOMED CT Content Request System
UTS	UMLS Terminology Services
XML	Extensible Markup Language

10. Special Topics

10.1 Introduction

As described in the Measure Development section, it is intended that contractors use a standardized process when developing measures. Developers should follow this process as closely as possible because it is designed to systematically guide the development of high caliber measures suitable for National Quality Forum (NQF) submission and the Centers for Medicare & Medicaid Services (CMS) implementation. However, some categories of measures and topics require specific approaches that diverge from the standard procedure.

This section will cover four topics:

- Cost and resource use measures are different from measures that evaluate the quality of care. NQF has adapted the standard evaluation criteria for these measures and has provided guidance for their development. This guidance is summarized in this section.
- Composite measures consist of two or more measures possibly already specified and endorsed. Measure development is unique for composites because the intended use of the composite and relationships between the component measures should be examined and understood.
- Outcome measures can be used in any setting and for all types of care, but design and use of those measures require unique consideration. This section will describe the special considerations that should be applied to outcome measures.
- Multiple chronic conditions measures address the challenge of evaluating care and outcomes for persons who have "two or more concurrent chronic conditions that collectively have an adverse effect on health status, function, or quality of life and that require complex healthcare management, decision-making, or coordination."[1] Specific considerations for measure development in these circumstances will also be described.

This section will describe how to approach development for these measure types where it varies from the general guide. There may be situations not covered in this section where a developer may require additional guidance. For those situations, contact the Measures Management team and the appropriate Contracting Officer Representative/Government Task Leader (COR/GTL).

10.2 Cost and Resource Use Measures

Background

The National Strategy for Quality Improvement in Health Care (National Quality Strategy) was called for under the Affordable Care Act and established national aims and priorities to guide local, state, and national efforts to improve the quality of health care in the United States.

[1]National Quality Forum. *Multiple Chronic Conditions Measurement Framework.* 2012. Available at: http://www.qualityforum.org/Projects/Multiple_Chronic_Conditions_Measurement_Framework.aspx. Accessed June 15, 2012.

The National Quality Strategy promotes quality health care that is focused on the needs of patients, families, and communities. The strategy presents three aims for the health care system:

- Better Care—Improve the overall quality, by making health care more patient-centered, reliable, accessible, and safe.
- Healthy People and Communities—Improve the health of the U.S. population by supporting proven interventions to address behavioral, social, and environmental determinants of health in addition to delivering higher-quality care.
- Affordable Care—Reduce the cost of quality health care for individuals, families, employers, and government.

Measures of cost and resource use can be used to assess the variability of the cost of healthcare with the objective being used as tools to direct efforts to make health care more affordable. Some terms related to measures addressing care that is affordable include:

- Cost of care—These are measures of the total health care spending, including total resource use and unit price(s), by payer or consumer, for a health care service or group of health care services, associated with a specified patient population, time period, and unit(s) of clinical accountability.
- Resource use—These measures are broadly applicable and comparable measures of health services counts (in terms of units or dollars) applied to a population or event (broadly defined to include diagnoses, procedures, or encounters). A resource use measure counts the frequency of defined health system resources; some may further apply a dollar amount (e.g., allowable charges, paid amounts, or standardized prices) to each unit of resource use—that is, *monetize* the health service or resource use units.
- Quality of care—Quality measures assess performance on the six Institute of Medicine (IOM) specified health care aims: safety, timeliness, effectiveness, efficiency, equity, and patient-centeredness.
- Efficiency—This term is associated with a measure of cost of care associated with a specified level of quality of care. [2]
- Value of care—This type of measure includes a specified stakeholder's preference-weighted assessment of a particular combination of quality and cost of care performance. A stakeholder can be an individual patient, consumer organization, payer, provider, government, or society, for example.

In a Call for Cost and Resource Use Measures, NQF describes the relationship of cost and resource use measures to efficiency:

> "Although resource use measures alone do not capture efficiency, they can be used as a building block toward understanding efficiency. Through the measurement of resource use, providers and

[2]National Quality Forum (NQF). *Measurement Framework: Evaluating Efficiency Across Patient-Focused Episodes of Care.* Washington, DC: NQF; 2009.

other stakeholders can associate a measure of cost with a specified level of quality of care toward understanding the efficiency of care for a population."[3]

Resource use measures can be developed for different units of analysis. The procedure described below is applicable to any of the following:

- Per capita-population and per capita-patient
- Per episode
- Per admission
- Per procedure

Procedure

Below is a summary of the steps described in the National Quality Forum's *Understanding & Evaluating Resource Use Measures (Phase I)* white paper available at: http://www.qualityforum.org/projects/efficiency_resource_use_1.aspx#t=1&s=&p=&e=1

Resource use measurement approaches can be viewed as having five main analytic functions:

- Data protocol
- Measure clinical logic
- Measure construction logic
- Adjustments for comparability
- Measure reporting

Step 1: Develop data protocol

Data input

Explicitly identify the types of data and aggregate or link these data so that the measure can be calculated reliably and validly. Merging data from different sources or systems must be done carefully so that errors in assumptions are not made. Some potential areas where problems may occur include:

- Difficulty in determining which data represent duplicates
- Different units of measurement used by the different data sources
- Different quality controls used by data sources

Data cleaning

- Check data to identify inaccurate, incomplete, or unreasonable data
- Correct data errors or omissions

Defining inclusion and exclusion criteria

Define the population to be included in the measure before defining the clinical logic.

[3]National Quality Forum. *Call for Measures: Cost and Resource Use*, January 2011.

Examples:

- Enrollment criteria may be established to only include claims for patients who were enrolled in a health plan for a full year.
- Only include patients who saw the physician at least once during the measurement period.

Step 2: Define measure clinical logic

Measures are usually identified as resource use measures for *acute conditions, chronic conditions,* or *preventive services*, which often affects the clinical logic. The analytic steps are designed to create appropriately homogeneous units for measurement.

Step 3: Define measure construction logic

Temporal

Decisions about when to start or end a measurement period must be specified for each measure. These time frames may be identified through clinical or evidence-based guidelines, expert opinion, or empirical data.

Assigning and triaging claims

Some examples of decisions that need to be addressed in managing claims data include:

- How to use different claims that provide information for the same event (especially those that result in an inflation of resource use amounts).
- When and how to map or feed claims from different sources into the same measure.
- When and which services trump other services.

Identifying units of resource use

The units of health services or resource use units must be identified and defined. Measure specifications measures must define and provide clear and detailed instructions on how to identify a single health-service unit, including the relevant codes, modifiers, or approaches to identify the amount.

Step 4: Adjusting for comparability

Define risk adjustment approach

Risk adjustment is designed to reduce any negative or positive consequences associated with caring for patients of higher or lower health risk or propensity to require health services. Resource use measures, including episode-based measures, generally risk adjust as part of the steps to address differences in patient characteristics and disease severity or stage.

Define stratification approach

Another type of adjustment is stratification, which is important where known disparities exist or where there is a need to expose differences in results so that stakeholders can take appropriate action. In

addition to exposing disparities, a measure may specify stratification of results within in a major clinical category (e.g., diabetes) by severity or other clinical differences.

Define costing methodology

The following costing methods may be used and will depend on the intended perspective.

- The count of services
- The actual amount paid
- Standardized prices

Step 5: Describe measure reporting

Attributing resource use measures

Resource measures are used to attribute the care provided as part of an episode of illness, the care of a population or event to a provider (e.g., physician, physician groups) or other entity (e.g., health plan) and in combination with quality or health outcome performance. It is easier to identify the appropriate provider to attribute the measure when the topic is narrowly defined, such as for a particular procedure. Measures for an episode of care or per capita measures are broader and often involve multiple providers, making valid attribution more difficult.

Care can be attributed to a single provider or multiple providers. Single attribution is designed to identify the decision maker, perhaps the primary care physician, and hold this individual responsible for all care rendered. Multiple attribution acknowledges that the decision maker, if there is one, has incomplete control over treatment by other physicians or specialists, even if the decision maker referred the patient to those other physicians.

Peer group identification and assignment

Unlike quality measures, which normally compare performance to an agreed-upon standard (e.g., providing flu vaccinations to a percentage of eligible patients) and direction for improvement (higher or lower performance is better), preferred resource use amounts often are not standardized; and it is not always clear if higher or lower resource use is preferable. Instead, resource use measure users often compare a physician's or entity's performance to the average performance of their peers. For this reason, it is essential to identify an appropriate peer group for comparison.

Calculating comparisons

- Observed-to-expected (O/E) ratio compares the value for each resource use measure attributed to a physician or entity (observed amount) and divides it by the average resource use within the identified peer group (expected amount—i.e., the amount of resource use expected if the entity measured were performing at the mean).
- More sophisticated statistical approaches, such as multi-level regression and Monte Carlo simulation, are also used.

Setting thresholds

Following the estimation of the value of a resource use measure, users must determine whether to apply thresholds or remove outliers to provide more context for the values. Outliers can be the result of inappropriate treatment, rare or extremely complicated cases, or coding error. Users often do not completely discard outliers, but rather examine them separately.

Providing detailed feedback

After all of the analytic steps are completed, users of resource use measures must decide which analytic results to include in any feedback or public reports.

Reporting with descriptive statistics

Decisions about which statistics must accompany the resource use measure results are critical. Factors influencing these include whether the results will be used for feedback or public reporting.

These types of analytic results can provide the detailed information necessary to make feedback actionable for all stakeholders. However, a number of options will need to be considered to provide reports with maximum actionability without information overload.

Step 6: Evaluate the measure

Following the processes described in the Information Gathering, Technical Expert Panel, Measure Testing and Public Comment sections, the measure contractor gathers the necessary information to evaluate the measure. The Measure Evaluation section provides a set of evaluation criteria, based on those of NQF, for the evaluation of Cost and Resource Use measures. The measure contractor should critically evaluate the measure. If the measure is to be submitted to NQF for endorsement the measure contractor should identify and document and potential weaknesses in the measures and suggest ways if any, to strengthen the measures. An assessment of the costs or time required to improve the measure should be presented along with the risks of not doing so.

Step 7: Document the measure in appropriate Measure Information Form (MIF)

Documentation of a Cost and Resource Use measure differs from those for a quality measures. The MIF can be modified to meet the needs of this type of measure. If directed by the COR and the measure is to be submitted for NQF endorsement, documentation in the MIF should use fields consistent with those used in the NQF submission form. When using the NQF online submission tool, select “Resource Use Measure Submission Form”.

10.3 Composite Measures

Background

NQF defines a composite measure as a combination of two or more individual measures in a single measure that results in a single score. These measures are useful for a variety of purposes. Composite measures can simplify and summarize a large number of measures or indicators into a more succinct piece of information. Stakeholders can then track broader ranges of information without getting

overwhelmed by data elements. Composite measures can be useful in situations such as public reporting Web sites and pay-for-performance programs. They take several components and translate them into a single metric summarizing overall performance. Composite measures can also be referred to as a composite index, composite indicator, summary score, summary index, or scale. Composite measures can evaluate various levels of the health care system, such as individual patient data, individual practitioners, practice groups, hospitals, or health care plans.

Procedure

Step 1: Identify the purpose

In the development of a composite measure, the first step is to identify the purpose for which the measure will be used (e.g., comprehensive assessment of adult cardiac surgery quality of care) and delineate the quality construct to be measured (e.g., four domains of cardiac surgery quality include perioperative medical care, operative care, operative mortality, and postoperative morbidity).[4]

The development of composite measures should follow these principles:[5]

- The purpose, intended audience, and scope of a composite measure should be explicitly stated.
- The individual measures used to create a composite measure should be evidence-based and reliable.
- The methodology used for weighting and combining individual measures into a composite performance measure should be transparent and empirically tested.
- The scientific properties of these measures, including reliability, accuracy, and predictive validity, should be demonstrated.
- Composites should be useful for clinicians and/or payers to identify areas for quality improvement.[6]

Step 2: Select component performance measures

Below are some considerations for the selection of measures that are to be included in the composite:

- In use or tested; proven to be reliable.
- Endorsed measures are preferred; however, a component measure might not be important enough in its own right as an individual measure, but it could be determined to be an important component of a composite.
- Have a plausible relationship (i.e., face validity) to the underlying construct.
- If available, consider using measures from an existing measure set.

---

[4] *Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety: Composite Measures,* National Quality Forum, June 2009.

[5] Peterson ED, Delong ER, Masoudi FA, et al. ACCF/AHA 2010 Position Statement on Composite Measures for Healthcare Performance Assessment: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures (Writing Committee to develop a position statement on composite measures). *Circulation.* 121(15):1780-1791.

[6] The Physician Consortium for Performance Improvement® Convened by the American Medical Association Measures Development, Methodology, and Oversight Advisory Committee: Recommendations to PCPI Work Groups on Composite Measures Approved by the PCPI in December, 2010.

Internal consistency may be assessed. However, it may have less relevance if the goal of the composite is to combine multiple distinct dimensions of quality as opposed to a single dimension. When such dissimilar elements are grouped together into a composite, the ability to evaluate such composites based on standard psychometric criteria is limited.

Example:

Timely reperfusion and use of secondary prevention discharge medications are both essential components of high-quality care for myocardial infarction (MI). Empirical studies of internal consistency may show a weak or absent association between performances on these two domains. However, if the objective of the composite measure is to provide a complete assessment of a hospital's performance on MI care processes, both measures may be appropriate for the composite.[7]

Step 3: Select a methodology and develop a scoring algorithm

- Ensure that the weighting and scoring of the components supports the goal that is articulated for the measure.
- Using a specified method, combine the component scores, into one composite.

Descriptions of the methodology of five common types of composite measure scoring are provided in Table 10-1. A listing of some of the advantages and disadvantages of each type, and examples of measures in the category are included.[8] The five types discussed are:

- All-or-none
- Any-or-none
- Linear combinations
- Regression-based composite measures
- Opportunity scoring

Table 10-1 Types of Composite Measure Scoring

Type of scoring	Advantages	Disadvantages	Examples/Evidence
All-or-None (defect-free scoring) **Process Measures** The patient is the unit of analysis. Only those patients who received all	◆ Promotes a high standard of excellence. ◆ Patient centric. ◆ Fosters a system perspective. ◆ Offers a more sensitive scale for assessing	◆ May waste valuable information. ◆ May weight common but less important processes more heavily than infrequent but important processes.	◆ Minnesota Community Measurement Optimal Diabetes Care measure. ◆ QIO 8th SOW Appropriate Care Measure (ACM) used hospital AMI, HF and PN process measures.[9]

[7]Peterson, et al. 2010.

[8]Much of the information in the table was obtained from Peterson et al. (2010).

[9]ACM Calculation: Add the total number of patient ACM numerators and divide by the total number of patient ACM denominators across the three topics (AMI, HF, PN) to calculate the ACM percentage. Example: a hospital has four cases submitted to the QIO clinical warehouse. Case #1 is an AMI patient that is qualified (eligible) for 3 of the 5 AMI measures and passes 2 of them. Case #2 is a HF patient that is qualified (eligible) for both of the 2 HF measures and passes both of them. Case #3 is a PN patient that is qualified (eligible) for 2 of the 3 PN measures and passes both of them. Case #4 is a HF patient that is not qualified for either of the 2 HF measures. For the ACM calculation, cases #1, #2, and #3 are in the denominator. Cases #2 and #3 are in the ACM numerator. The hospital's ACM rate is 2/3 or 66.7%.

Type of scoring	Advantages	Disadvantages	Examples/Evidence
indicated processes of care are counted as *successes*. Performance is defined by the proportion of patients receiving all of the specified care processes for which they were eligible. No credit is given for patients who receive some but not all required items.	improvements. ◆ Especially useful for those conditions in which achieving a desired clinical outcome empirically requires reliable completion of a full set of tasks (that is, when partial completion does not gain partial benefit).	◆ The provider who achieved 4 of 5 measures appears the same as the provider who achieved none of 5 measures. ◆ The all-or-none approach will amplify errors of measurement (one unreliable component measure will contaminate the whole score) so that it is essential that each of the component measures be well designed.	◆ IHI Bundles: ventilator, central line. ◆ STS Perioperative Medical Care, a process bundle of 4 medications (preoperative beta blockade and discharge anti-platelet, beta blockade, and lipid-lowering agents). ◆ Study using Premier SCIP data; adherence measured through a global all-or-none composite infection-prevention score was associated with a lower probability of developing a postoperative infection. However, adherence reported on individual SCIP measures was not associated with a significantly lower probability of infection.[10]
Any-or-None Outcome Measures Similar to all-or-none, but is used for events that should not occur. The patient is the unit of analysis. A patient is counted as *failing* if he or she experiences at least 1 adverse outcome from a list of 2 or more adverse outcomes.	◆ Promotes a high standard of excellence. ◆ Useful when component measures are rare events.	◆ Particularly problematic when rare but important outcomes are mixed with common but relatively unimportant outcomes, because the composite is likely to be dominated by the outcome that occurs most frequently.	◆ STS Postoperative Risk-Adjusted Major Morbidity, any of the following—renal failure; deep sternal wound infection; re-exploration; stroke; and prolonged ventilation/intubation. This is an "any or none" measure, requiring the absence of all such complications.
Linear Combinations Can be simple average or weighted average of individual measure scores.	◆ Has the advantage of simplicity and transparency.	◆ Does not account for potential differences in the validity, reliability, and importance of the different individual measures. ◆ Equal weighting may be undesirable if there is a considerable imbalance	◆ Premier/CMS Hospital Quality Incentive Demonstration uses a composite of process and outcome to measure quality for CABG. The composite quality score (CQS) was based on an equally weighted combination of 7

[10] Stulberg JJ, Delaney CP, Neuhauser DV, Aron DC, Fu P, Koroukian SM. Adherence to Surgical Care Improvement Project Measures and the Association With Postoperative Infections. *JAMA*.303(24):2479-2485.

Type of scoring	Advantages	Disadvantages	Examples/Evidence
		in the numbers of measures from different domains. ◆ Different stakeholders have different priorities; one weighting method may not meet the needs of all potential users. ◆ When items with a small standard deviation are averaged with items with a large deviation, items with the large standard deviation tend to dominate the average. ◆ If items are combined that are not positively or negatively correlated with one another (i.e., covary), the resulting composite score may not possess reasonable properties to allow meaningful differentiation among patients, and may not measure a single construct. This issue can be mitigated by pursuing latent factor analysis strategies to ensure that items cohere to form a reasonable single score for a construct.	measures (4 process measures and 3 outcome measures). The actual publicly reported data suggest that the CQS was more heavily influenced by process measures than would have been expected by the apparent 4:3 weighting. ◆ The US News & World Report Index of Hospital Quality for heart and heart surgery is a linear combination of 3 equally weighted components: reputation, risk-adjusted mortality, and structure. Although the 3 components are weighted equally, a hospital's reputation score has the highest correlation with its overall score, in comparison; the Mortality Index appears to have much less influence. ◆ The AHRQ PSI composite measure uses a weighted average of various individual component measures. The weighting was determined by an expert panel.
Regression-Based Composite Measures If a certain outcome is regarded as a gold standard, the weighting of individual items may be determined empirically by optimizing the predictability of the gold standard end point.	◆ The weight assigned to each item is directly related to its reliability and the strength of its association with the gold standard end point. ◆ Regression-based weighting may be appropriate for predicting specific end points of interest	◆ Weighting may not be optimal for objectives, such as motivating health care professionals to adhere to specific treatment guidelines.	◆ Leapfrog developed surgical "survival predictor" composite measures to forecast hospital performance, based on prior hospital volumes and prior mortality rates. An empirical Bayesian approach was used to combine mortality rates with information on hospital volume at each hospital. The observed mortality rate is

Type of scoring	Advantages	Disadvantages	Examples/Evidence
			weighted according to how reliably it is estimated, with the remaining weight placed on hospital volume.
Opportunity Scoring Opportunity scoring counts the number of times a given care process was actually performed (numerator) divided by the number of chances a provider had to give this care correctly (denominator). Unlike simple averaging, each item is implicitly weighted in proportion to the percentage of eligible patients, which may vary from provider to provider.	◆ Provides an alternative to simple averaging often used for aggregating individual process measures. ◆ Has the advantage of increasing the number of observations per unit of measurement, consequently potentially increasing the stability of a composite estimate, particularly when the sample size for individual measures is not adequate.	◆ Rate is influenced by the most common care processes, regardless of whether or not they are the most important methods.	◆ The opportunity model was developed for the Hospital Core Performance Measurement Project for the Rhode Island Public Reporting Program for Health Care Services in 1998. ◆ CMS/Premier Hospital Quality Incentive (HQI) Demonstration project uses the opportunity scoring method for the process composite rate for each of 5 clinical areas. The sum of all the numerators is divided by the sum of all the denominators in each clinical area.

Step 4: Evaluate the composite

Even though all of the component measures must individually meet evaluation criteria, the composite measure as a whole also must meet evaluation criteria. (Refer to the Measure Evaluation Section, Special Considerations-Composites)

A Composite Measure Evaluation Tool is available to aid in the application of the criteria.

Step 5: Document technical specifications of the composite

Similarly, though technical specifications of all components of the composite may already be documented, they should also be completed for the composite. The Measure Information Form and Measure Justification forms are aligned with the requirements of the NQF Measure Submission Form and will guide the measure developer to ensure that the technical specifications are sufficient and complete. Composite measure technical specifications should be included with other measure documentation forms for submission to the COR/GTL for approval.

10.4 Outcome Measures

Background

Outcome measures are one of the three basic types of measures used to assess the quality of health care: structure, process, and outcome. Outcome measures assess results of health care experienced by

patients: patients' clinical events, patients' recovery and health status, patients' experiences in the health system, and efficiency/cost. Outcomes are dependent upon process of care because they are by definition the results of the actions of the health care system.

Though multiple provisions in the ACA support the development of quality measures, and Section 10303 specifically calls for outcome measures to be developed. Outcomes capture what patients and society care most about—the results of care. However, outcome measures are often difficult to construct in a way that produces valid and meaningful information. Section 10303 identifies three technical challenges when developing outcome measures: risk adjustment, sample size, and accountability. Intermediate outcome measures are measures that aim to meet specific thresholds or other results of clinical care that have been shown to affect the desired health outcome (positively OR adversely). Most outcome measures, other than those which measure intermediate outcomes, require risk adjustment.

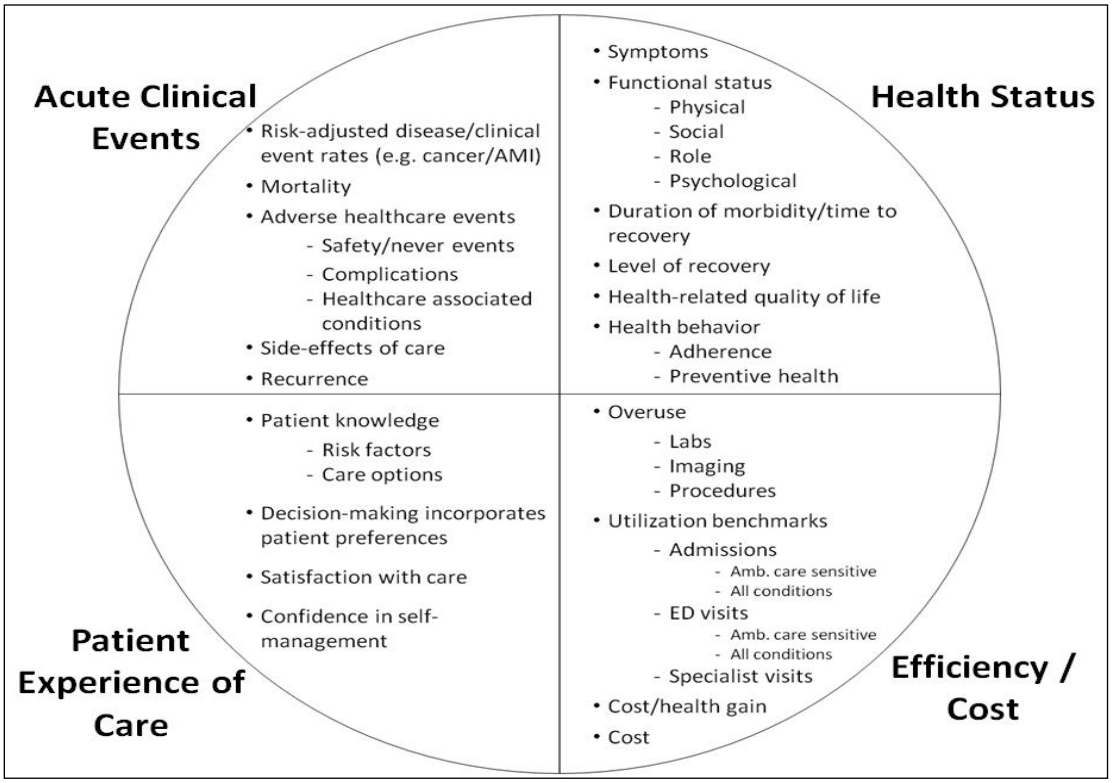
The subject of risk adjustment is described in more detail in the Risk Adjustment section.

CMS Outcomes Quality Compass

CMS recently commissioned a white paper to inform the agency in selecting outcome measures to meet the requirements of Section 10303 of the ACA. In this white paper, four categories of potential outcomes are identified: Patient Acute Clinical Events, Patient Health Status, Patient Experience of Care, and Cost/Efficiency. CMS presents these categories in an "Outcomes Quality Compass" which provides a menu of outcome measures (Figure 10-1).[11]

[11]Centers for Medicare and Medicaid Services. *Selecting Outcomes Measures for Section 10303 of the Patient Protection and Affordable Care Act: A White Paper.* 2011.

Figure 10-1 Outcomes Quality Compass



Patient self-reported outcomes

Outcomes can be measured in a variety of ways. In addition to the usual methods (e.g., administrative data, patient assessment instruments, and medical records), patient self-reported data provides a rich data source for outcomes. Some examples of patient self-reported data include:

- Patient Reported Outcomes Measurement Information System (PROMIS)—Funded by the National Institutes of Health (NIH), is a system of highly reliable, valid, flexible, precise, and responsive assessment tool that measures patient self–reported health status.
- Health Outcomes Survey (HOS)—The Medicare HOS is the first outcomes measure used in Medicare Advantage plans. The goals of the Medicare HOS program are to gather valid and reliable health status data in Medicare managed care for use in quality improvement activities, plan accountability, public reporting, and health improvement. All managed care plans with Medicare Advantage (MA) contracts must participate.
- Focus On Therapeutic Outcomes, Inc. (FOTO)—Is used to measure the functional status of patients who received outpatient rehabilitation through the use of self-reported health status questionnaires. Because the measures are taken at intake, during, and at discharge from rehabilitation, the change in functional status can be assessed.

Procedure

Step 1: Choose and define an outcome

The outcome must be appropriate, the definition of the outcome must clearly define what is counted and is not counted, and one must have the capability to collect reliable and valid outcome data. The selected outcome, even if actually a surrogate to the primary outcome of interest, must provide an adequate number of events for valid statistical modeling. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.[12]

An appropriate outcome has clinical or policy relevance. For example, mortality would not be an appropriate measure for cataract surgery as it is extremely rare and nearly irrelevant. A better outcome measure in this instance might be a clinically meaningful measure of improvement in vision. Whenever possible, clinical experts should be consulted to more relevantly define appropriate and meaningful outcomes. If a technical expert panel (TEP) is to be used, refer to the Technical Expert Panel section for the standardized processes.

Step 2: Determine the appropriate time frame

The time frame for the outcome must be meaningful. The evaluation of the outcome must also be based on a standardized period of assessment.

If the period of the outcome assessment is not standardized, such as the assessment of an event during hospitalization, the evaluation may be biased because health care providers have different practice patterns (e.g., varying lengths of stay).

A study on hospital mortality was conducted at a time when hospitals were shortening patient lengths of stay. The study found that in-hospital mortality had decreased, but 30-day mortality had not. Instead of dying in hospitals, patients were dying in the other settings they were transferred to after their hospitalization.[13] This example illustrates how the standardized time frame of 30 days would have made the measure more valid than only using the inpatient mortality rates.

The time frame should also be standardized so that sufficient time is allowed to reliably capture the relevant outcome. If the timing of the outcome measurement is too long after care is provided, it will be more difficult to link the quality of care from that provider to that outcome.

Example:

Delaying heart failure readmission rates for 180 days after hospital discharge would make it difficult to attribute the readmission to the discharging facility. Many possible intervening events may confound the attribution.

[12]National Quality Forum. *Measure Evaluation Criteria*. Importance Criterion, Note 3. January 2011.
[13]Baker DW, Einstadter D, Thomas CL, et al. Mortality Trends During a Program that Publicly Reported Hospital Performance. *Medical Care.* 2002a. 40 (10): 879-90.

Step 3: Determine other key issues: sample size and attribution

While some measure concepts have face validity and intuitively appear to be useful, developers must gauge what constitutes an adequate sample size. For example, it is often difficult to obtain an adequate sample size for individual physician-level outcome measures because adequate risk adjustment requires a larger sample. Reporting at the group-level may address this concern. The Society of Thoracic Surgeons (STS), for example, reports at the group-level rather than physician-level.

The measure contractor must evaluate which providers or levels of analysis are appropriate to allow attributing responsibility for the outcome. The outcome should be under control of the provider being measured. In order to promote collaboration and accelerate improvement in the health care environment, it may be appropriate to hold providers accountable for outcomes in which they share responsibility with other providers.

Step 4: Determine appropriate risk adjustment

As a rule, intermediate outcome measures do not require risk adjustment. Stratification of results by subpopulation may be more appropriate when rate variability exists for these subpopulations. This should be done in order to identify and reduce health care disparities. Refer to the Risk Adjustment section for discussion on determining the need for risk adjustment and development, and evaluation of risk adjustment models.

Step 5: Document measure technical specifications

Develop detailed and precise measure specifications using the Measure Information Form following the guidance provided in the Technical Specifications section.

10.5 Multiple Chronic Conditions Measures

Much of the measures development process described throughout this Blueprint apply directly to measures development for Multiple Chronic Conditions, but some differences exist. They will be described here.

Background

The Robert Wood Johnson Foundation reported that in 2006 more than one in four Americans had multiple chronic conditions.[14] These individuals constitute a particular challenge to the health care system because their conditions complicate each other, are ongoing, and are very costly to both the persons involved and the nation overall. The effects of their co-morbidities are more than simply additive. They multiply both morbidity and mortality.[15] In 2011 CMS found that Medicare beneficiaries with multiple conditions were the heaviest users of health care services.[16] For those with six or more

[14]Anderson G. *Chronic care: making the case for ongoing care.* Princeton, NJ: Robert Wood Johnson Foundation, 2010. http://www.rwjf.org/files/research/50968chronic.care.chartbook.ppt. Last accessed June 21, 2012.
[15]Tinetti ME, McAvay GJ, Chang SS, et al. Contribution of multiple chronic conditions to universal health outcomes. *Journal of the American Geriatric Society.* 2011; 59(9):1686-91.
[16]Centers for Medicare & Medicaid Services. *Chronic Conditions among Medicare Beneficiaries, Chart Book.* Baltimore, MD. 2011.

chronic conditions, two thirds were hospitalized and they accounted for about one half of Medicare spending on hospitalizations.[17] However, very few measures exist that are specifically designed to evaluate the quality of care provided to these people.[18]

Multiple chronic conditions (MCC) definition

HHS recently contracted with NQF to develop a measurement framework for persons with MCCs. The NQF *Multiple Chronic Conditions Measurement Framework* defined MCC as follows:

- Persons with MCCs are defined as having two or more concurrent chronic conditions that collectively have an adverse effect on health status, function, or quality of life and that require complex healthcare management, decision-making, or coordination.
- Assessment of the quality of care provided to the MCCs population should consider persons with two or more concurrent chronic conditions that require ongoing clinical, behavioral, or developmental care from members of the healthcare team and act together to significantly increase the complexity of management and coordination of care—including but not limited to potential interactions between conditions and treatments.
- Importantly, from an individual's perspective the presence of MCCs would:
  - Affect functional roles and health outcomes across the lifespan.
  - Compromise life expectancy.
  - Hinder a person's ability to self-manage or a family or caregiver's capacity to assist in that individual's care.[19]

Need for measure development

Though persons with MCCs represent a growing proportion of society who use an increasingly large amount of health care services, existing quality measures do not adequately address their needs.[20] Current quality measures are largely based on performance standards derived from clinical practice guidelines for management of specific diseases.[21] Patients with multiple conditions have often been excluded from the evidence generating clinical trials that form the basis of many clinical practice guidelines. The randomized clinical trials used in clinical practice guideline development focus mainly on single diseases to produce robust guidance for specific disease treatments. Rigid adherence to these disease specific guidelines could potentially harm those with MCCs. For example medications prescribed in adherence to guidelines for several diseases individually may result in a patient suffering adverse effects of polypharmacy.[22] Few measures exist to evaluate inappropriate care in these

[17]Ibid.
[18]National Quality Forum. *Multiple Chronic Conditions Measurement Framework.* 2012. Available at: http://www.qualityforum.org/Projects/Multiple_Chronic_Conditions_Measurement_Framework.aspx. Accessed June 15, 2012.
[19]Ibid. p. 7.
[20]National Quality Forum. *Multiple Chronic Conditions Measurement Framework.* 2012. Available at: http://www.qualityforum.org/Projects/Multiple_Chronic_Conditions_Measurement_Framework.aspx. Accessed June 15, 2012.
[21]Tinetti ME, Bogardus ST, Agostini JV. Potential pitfalls of disease-specific guidelines for patients with multiple conditions. *New England Journal of Medicine.* 2004;351:2870–4.
[22]Ibid.

situations. "An in-depth consideration of these complex issues is important to address measurement for individuals with MCCs adequately."[23]

Considerations for measure development targeting persons with MCCs

During step 1—when choosing appropriate measure concepts consider

Without evidence based guidelines specifically directed to care of persons with multiple chronic conditions, best practices may remain up to the clinical judgment of the providers. However, measurable quality topics do exist that are especially pertinent to people with multiple chronic conditions. The following measurement concepts were identified as having potential for high-leverage in quality improvement for patients with MCCs.[24]

- Optimizing function, maintaining function, or preventing further decline in function
- Seamless transitions between multiple providers and sites of care
- Patient important outcomes (includes patient-reported outcomes and relevant disease-specific outcomes)
- Avoiding inappropriate, non-beneficial care, including at the end of life
- Access to a usual source of care
- Transparency of cost (total cost)
- Shared accountability across patients, families, and providers
- Shared decision-making

These measure concepts represent cross-cutting areas with the greatest potential for reducing factors of cost, disease burden, and improving well-being that are highly valued by providers, patients and families.

During step 2—when determining how to address key issues

Guiding principles

The NQF Framework identified that quality measures for persons with Multiple Chronic Conditions should be guided by several principles. Quality measures should: [25]

- Promote collaborative care among providers.
- Incorporate several types of measures that address appropriateness of care.
- Prioritize optimum outcomes that are jointly established by considering patient preferences.
- Address shared decision-making.
- Assess care longitudinally, changes in care over time, such as delta measures of improvement or maintenance rather than attainment.
- Be as inclusive as possible, not excluding persons with multiple conditions.

---

[23]National Quality Forum. *Multiple Chronic Conditions Measurement Framework.* 2012. Available at: http://www.qualityforum.org/Projects/Multiple_Chronic_Conditions_Measurement_Framework.aspx. Accessed June 15, 2012.
[24]Ibid p. 9.
[25]National Quality Forum. *Multiple Chronic Conditions Measurement Framework.* 2012. Available at: http://www.qualityforum.org/Projects/Multiple_Chronic_Conditions_Measurement_Framework.aspx. Accessed June 15, 2012.

- Illuminate and track disparities through stratification and other approaches.
- Use risk adjustment for comparability (of outcome measures only) with caution as it may obscure serious gaps.
- Standardize inputs from multiple sources, particularly patient reported data.

Time frame issues to consider

Measurement time frame is particularly important with chronic conditions because the very nature of chronic conditions requires observation over time. Especially in the case of outcome measures for beneficiaries with multiple conditions, it is very difficult to know where to attribute responsibility unless the measurement time frame is carefully considered and specified. Measures for this population should assess care across episodes, across providers and staffing, using a longitudinal approach. Delta measures of improvement (or maintenance rather than decline) over extended periods of time are particularly relevant in this population.

Attribution issues to consider

Issues of attribution are compounded when adding the factor of multiple chronic conditions. Since the multiple conditions also mean multiple providers, it becomes difficult to choose who should be credited for good outcomes and which provider gave inadequate care when the treatment for one condition might exacerbate the other. These issues may require a more aggregated level of analysis such as at a provider group level or population rather than individual level. Since beneficiaries with multiple chronic conditions see multiple providers, it would be more appropriate to measure and attribute the outcomes for the population to the care provided by the team of providers.

Methodological issues to consider

Quality measures for this population should be designed to be as inclusive as possible. Instead of excluding these individuals from the denominator of measures, methodological approaches should be designed to reveal and track variances in care and outcomes.

The empirical link between quality processes with the outcomes of those health care processes is even more difficult to establish when dealing with multiple and chronic conditions. Risk adjustment should be used with caution in the situation of multiple chronic conditions. Stratification may allow quality comparison across populations without masking important distinctions of access, care coordination and other issues. Refer to the Risk Adjustment section for in-depth discussion on how to determine when risk adjustment is appropriate and how to evaluate risk adjustment models when they are applied.

Quality measures for this population should address quality across multiple domains. Measures should be harmonized across level of the health care system to provide a comprehensive picture of care.

Data gathering issues to consider

There may be difficulties gathering data systematically, especially for this population. Particularly, patient-reported data may be difficult to collect because of the interacting conditions. Interpretation of different types of data is needed as the data may come from multiple providers, multiple sources, in

multiple types, and over extended periods of time. It is important for measure developers to standardize data collection methods. Health Information Technology may be leveraged for more reliable data collection.

During step 3—when testing and evaluating measures for persons with MCC

Evaluation methods described elsewhere in the Blueprint also apply to measures of quality care for persons with Multiple Chronic Conditions. In addition, the MCC measures should successfully carry out the guiding principles from the NQF framework listed above. Functional status and other outcomes should be examined using delta measures of change over time. If new tools and methods of data collection are developed, those tools must also be carefully assessed. Alpha testing may be particularly important in the formative phase not only for new tools designed for these types of measures but also to test the feasibility of linking data from a variety of sources.

During step 4—document measure technical specifications

There are no differences in this step. Technical specifications for these measures must meet the same criteria as any other measure or measure set. Develop detailed, precise measure specifications using the Measure Information Form following the guidance provided in the Technical Specification section.

11. Risk Adjustment

11.1. Introduction

There are many terms that describe the concept of risk adjustment. Risk adjustment, severity adjustment, and case-mix adjustment are all often used to describe similar methods. This section of the Blueprint references evidence-based risk-adjustment strategies that encompass both statistical risk models and risk stratification using language employed by the National Quality Forum (NQF). As part of a risk-adjustment strategy, NQF recommends use of risk models in conjunction with risk stratification when use of a risk model alone would result in obscuring important healthcare disparities. In this section, the term risk adjustment refers to the statistical process used to identify and adjust for differences in population characteristics (i.e., risk factors) before comparing outcomes of care. The words risk adjusted outcome, risk factor, and risk adjustor are used to describe expressions related to a risk adjustment model. In contrast, the term risk stratification, as used in this section, refers to the separate reporting of different groupings of data that may or may not be adjusted by a risk model.

Within this framework of a risk adjustment strategy, the purpose of any measure risk adjustment model is to facilitate fair and accurate comparisons of outcomes across health care organizations, providers, or other groups.[1] Risk adjustment of health care measures is encouraged because the existence of risk factors before or during health care encounters may contribute to different outcomes regardless of the quality of care received, and adjusting for these risk factors can avoid misleading comparisons. However, risk adjustment models for publicly reported quality measures should not obscure disparities in care associated with race, socioeconomic status, or gender. The exploration of a risk adjustment strategy (i.e. the use of a statistical risk adjustment model and, if necessary, risk stratification for selected populations[2]) is required for measures developed using the Blueprint. For a measure to be accepted by the Centers for Medicare & Medicaid Services (CMS) and endorsed by NQF, the measure developer must demonstrate the appropriate use of a risk adjustment strategy. Rationale and strong evidence must be provided if a risk adjustment model or risk stratification is not used. Consequently, it is the measure developer's responsibility to determine if variation in factors intrinsic to the patient should be accounted for before outcomes can be compared, and how to best apply these factors in the measure specifications. It is important to remember that risk adjustment does not itself provide the answers to study questions about measures, but instead, provides a method for determining the most accurate answers.[3] The purpose of this section is to provide guidance to CMS measure contractors regarding the nature and use of a risk adjustment model in quality measurement.

[1] An example of different methods to adjust within and across groups is found in The American Academy of Actuaries May 2010 Issue Brief titled "Risk Assessment and Risk Adjustment" that discusses risk adjustment in the context of the Patient Protection and Affordable Care Act (PPACA) and issues needing attention and accommodation prior to the 2014 inclusion of small group markets.

[2] NQF policies generally preclude the use of risk factors that obscure disparities in care associated with factors such as race, socioeconomic status, or gender.

[3] Boston, MA. Management Decision and Research Center; Washington, DC: *VA Health Services Research and Development Service in collaboration with Association for Health Services Research* 1997, W84.AA1 R595 1997.

11.2 Deliverables

- Risk Adjustment Methodology Report that includes full documentation of the risk adjustment model or rationale and data to support why no risk adjustment or stratification is needed.
- Measure Information Forms (MIFs) with completed risk adjustment sections for each measure.

11.3 Attributes of Risk Adjustment Models

The measure contractor must evaluate the need for a risk adjustment strategy (i.e., risk adjustment, stratification, or both) for all potential measures, and test the adequacy of any strategies used. In general, a risk adjustment model possesses certain attributes. Some of these attributes are listed in Table 11-1 Attributes of Risk Adjustmentwhich was partially derived from a description of preferred attributes of models used for publicly reported outcomes (Krumholz et al., 2006[4]). Each of these is described in detail in the sections below. Attributes for stratification models are described in Step 3 of Section 11.4—Procedure.

Table 11-1 Attributes of Risk Adjustment Models

Attribute	Description
Sample definition	Sample(s) should be clearly defined and clinically appropriate for the measure's risk adjustment
Appropriate time frames	Time frames for model variables should be clearly defined, sufficiently long to observe an outcome, and recent enough to retain clinical credibility
High data quality	The data should be reliable, valid, and complete
Appropriate variable selection	Selected adjustment or stratification variables should be clinically meaningful
Reasonable analytic approach	Analytic approach must be scientifically rigorous and defensible, and take into account multilevel organization of data (if necessary)
Complete documentation	Risk adjustment and/or stratification details and it's performance must be fully documented and all known issues disclosed

Sample definition

The sample(s) should be clearly and explicitly defined. Any criteria used to select the sample should be defined. Risk adjustment models will only function (i.e., generalize) to the extent that the samples used to develop, calibrate, and validate them appropriately represent the overall population. Samples are microcosms such that the distributions of characteristics and their interactions should mimic those in

[4]Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group. *Circulation* 2006;113, 456 – 62.

the overall population. Researchers need to explain their rationale for using selected samples and offer justification of the sample's appropriateness.

Appropriate time frames

All of the decisions regarding the selection of the time frame should be clearly stated and explained in the measure documentation. Time frames used to identify risk factors and outcomes should be clinically appropriate and clearly stated. Risk factors should be present at the start of care to avoid mistakenly adjusting for factors arising due to deficiencies in care being measured, and outcomes should occur soon enough after care to establish that they are the result of that care. For example, renal failure is one of the comorbidities that may be used for risk adjustment of a hospital mortality measure. If poor care received at the hospital caused the patient to develop renal failure *after* admission, it would be inappropriate to adjust for it for that patient.

The evaluation of outcomes must also be based on a standardized period of assessment. If the periods of the outcome assessments are not standardized, such as the assessment of events during hospitalization, the evaluation may be biased because health care providers have different practice patterns (e.g., varying lengths of stay). This potential issue was illustrated in a study of hospital mortality at a time when hospitals were shortening patient lengths of stay. Baker, Einstadter, Thomas et al. (2002)[5] found that in-hospital mortality had decreased, but 30-day mortality had not. Instead of dying in hospitals, patients were dying in other settings they were transferred to after their hospitalization. This example illustrates how the standardized time frame of 30 days would have made the measure more valid than using only the inpatient mortality rates, and would have allowed sufficient time to reliably capture the relevant outcome.

High data quality

The measure contractor must ensure that the data used for risk adjustment are of high quality. Considerations in determining the quality of data are as follows:

- The data were collected in a reliable way. That is, the method of collection must be reproducible with very little variation between one collection and another if the same population was the source.
- The data collected are as recent as possible. If the measure contractor were using 1990 data in a model designed to be used tomorrow, many people would argue that the health care system is different enough from 1990 that the model may not be relevant.
- The data collected are as complete as possible. The data should contain as few missing values as possible. Missing values are hard to interpret and lower the validity of the model.
- Documentation of the sources of the data including when it was collected, if and how it was cleaned and manipulated, and its assumed quality should be fully disclosed.

[5]Baker D, Einstadter D, Thomas C, et al. Mortality Trends During a Program that Publicly Reported Hospital Performance. *Medical Care*. 40;10:879-890.

Appropriate variable selection

The risk adjustment model variables should be clinically meaningful. When developing a risk-adjusted model, the clinical relevance of included variables should be apparent. This serves two purposes: It contributes to the face validity of the model, and it increases the likelihood that the model will explain variation already identified by health care professionals as important to the outcome. Less complex relationships between patient factors and the outcome are likely to have the highest face validity and be optimal for use in a model. The strength of the relationships required to retain adjustment factors ultimately depends upon the conceptual model, but rarely is a factor included in a model that is not substantively associated with the outcome variable.

Occasionally, less obvious variables may be included in the risk adjustment model based upon prior research. This situation may arise when direct assessment of a relevant variable is not possible, and the use of a substitute (i.e., proxy) variable is required. However, the relevance of these substitute variables should be clinically and empirically apparent. For example, medications taken might be useful as a proxy for illness severity or progression of a chronic illness, provided practice guidelines or prior studies clearly link the medication patterns to the illness severity or trajectory. Similarly, inclusion of variables previously shown to moderate the relationship between a risk adjustor and the measure may be included. Moderating variables are generally included as interaction terms in a model (e.g., a prior mental health diagnosis may be only weakly associated with a measured outcome, but it may interact with another variable to strongly predict the outcome). Interaction terms require specialized data coding and interpretation.

Reasonable analytic approach

An appropriate statistical model is determined by many factors. Logistic regression or hierarchical logistic regression is often used when the outcome is categorical; but, in certain instances, the same data may be used to develop a linear regression model when key statistical assumptions are not violated. Selecting the correct statistical model is absolutely imperative, because an incorrect model can lead to entirely erroneous results. The analytic approach should also take into account any multilevel organization of data, which is typically present when assessing institutions such as hospitals.

Risk factors retained in the model should account for substantive and significant variation in the outcome. Overall differences between adjusted and unadjusted outcomes should also be practically/clinically meaningful. Moreover, risk factors should not be related to the stratification factors. A statistician can provide the measure contractor team with the necessary guidance and recommendations as to the most useful variable formats and appropriate models.

Complete documentation

Transparency is one of the key design principles in the Blueprint. When researchers do not disclose all of the steps that were used to create a risk adjustment model, others cannot understand or fully evaluate the model. Recent Department of Health and Human Services policies emphasizes transparency. National Quality Forum (NQF) policy on the endorsement of proprietary measures promotes the full disclosure of all aspects of a risk adjustment model used in measure development.

The methods used, including the risk adjustment method, performance of the risk adjustment model and its components, algorithms, as well as the sources of the data and methods used to clean or manipulate the data should be fully described. Documentation should be sufficient to allow others to reproduce the findings. The development of the measure and its documentation is expected to incorporate statistical and methodological recommendations from a knowledgeable statistician.

11.4 Procedure

The following steps are recommended in the development of a risk adjustment model. Note that some models may not lend themselves appropriately to all of these steps, and an experienced statistician or clinical expert can determine the need for each step.

Step 1: Choose and define an outcome

While risk-adjustment should not be applied to structure and process measures that are entirely within the measured provider's control, risk-adjustment may be necessary for outcome measures that are not fully within the measured providers' control [6] (e.g., re-admission rates, mortality, and length of stay). When selecting outcomes that are appropriate for risk-adjustment, the time frame for the outcome must be meaningful, the definition of the outcome must clearly define what is counted and not counted, and one must be able to collect the outcome data reliably. An appropriate outcome has clinical or policy relevance. It should occur with sufficient frequency to allow statistical analysis, unless the outcome is a preventable and serious health care error that should never happen (i.e., a "Never Event"). Outcomes should be evaluated for both validity and reliability (Refer to the Measure Testing section). Whenever possible, clinical experts, such as those participating in the technical expert panel (TEP) should also be consulted to help define appropriate and meaningful outcomes.

Step 2: Define the conceptual model

A clinical hypothesis or conceptual model about how potential risk factors relate to the outcome should be developed *a priori*. The conceptual model serves as a map for the development of a risk adjustment model. It defines the understanding of the relationships behind the variables, and as such, will help identify which risk factors, patients, and outcomes are important, and which can be excluded. Because the cost of developing a risk adjustment model may be prohibitive if every potential risk factor is included, the conceptual model also enables the developer to prioritize among risk factors, and evaluate the cost and benefit of data collection. Alternatively, the existence of large databases and modern computing power allow for statistical routines (e.g., boot strapping, jack knifing, etc.) to explore the data for relationships between outcomes and potential adjustment factors that might not yet be clinically identified, but empirically exist.

[6] National Quality Forum. *National Voluntary Consensus Standards for Ambulatory Care—Measuring Healthcare Disparities: A Consensus Report.* Available at: http://www.qualityforum.org/Publications/2008/03/National_Voluntary_Consensus_Standards_for_Ambulatory_Care%E2%80%94Measuring_Healthcare_Disparities.aspx Accessed July 31, 2012.

The first step in developing or selecting the conceptual model is identifying the relationship among variables. This may include:

- Conducting a review of clinical literature and canvassing expert opinion to establish variable relationships.
- Obtaining expert opinion. The experts consulted should include health care providers with clinically relevant specialties. A TEP may be used if diverse input is sought. Refer to the Technical Expert Panel section for the standardized process used for convening a TEP.
- When appropriate data are available, automated computer routines can be used to identify potential factors for consideration by subject matter experts.

Step 3: Identify the risk factors and timing

Use of a conceptual model and clinical expertise promotes selection of risk factors with the following attributes:

- Clinically relevant
- Reliably collected
- Associated with the outcome
- Clearly defined
- Identified using appropriate time frames

In addition to these attributes, risk factors should also align with NQF policies for endorsed measures. These policies generally preclude the use of risk factors that obscure disparities in care associated with factors such as race, socioeconomic status, or gender. In the Table 11-2 Examples it provides a few examples of factors that generally should not be used in risk adjustment models for quality measures, even if their inclusion improves the predictive ability of the model. When such factors exist, development of measures that stratify the population may be more appropriate than their inclusion in a risk adjustment model.

Table 11-2 Examples

Risk Factor	Discussion
Race or ethnicity	Populations representing certain races/ethnicities are at different levels of risk for disease and mortality, and they have different level-of-care needs. Risk models should not obscure disparities in care for populations by including factors that are associated with differences or inequalities in care such as race or ethnicity. It is preferable to stratify by this factor, rather than use it in a risk adjustment model.
Medicaid status	While there have been CMS programs that warrant the use of Medicaid status in a risk adjustment model (e.g. Medicare Health Outcomes Survey or comparisons between groups with diverse proportions of dual-eligible beneficiaries), NQF policy discourages the use of risk factors that obscure differences associated with socioeconomic status. Consequently, it is preferable to stratify by this factor, or consult with CMS prior to including it in any competing or alternate risk adjustment model.

Risk Factor	Discussion
Sex	Males and females often show differences relative to treatment, utilization, risk levels for mortality and disease, and needed level of care. While restricting a measure to a single sex using exclusion criteria may be reasonable, inclusion of sex in a risk model should be avoided to prevent obscuring disparities in care. It is preferable to stratify by sex, rather than use it in a risk adjustment model.

Step 4: Acquire data (sample, if necessary)

Health care data can be acquired from many sources but the three most frequently used are administrative data, patient record data, and survey data. Of these, the most common source of data for developing risk adjustment models is administrative data reported by the provider. Once the data sources are acquired, relevant databases may need to be linked and various data preparation tasks performed, including an assessment of the data reliability if not previously confirmed. If one or more samples are to be used, they should be drawn using predefined criteria and methodologically sound sampling techniques. Testing to determine the suitability of data sources and testing for differences across data sources may also be necessary (refer to the alpha and beta testing discussion in the Measure Testing section).

Step 5: Model the data

In addition to the clinical judgment used to define the conceptual model and candidate variables, empirical modeling should also be conducted to help determine the risk factors to include or exclude. A number of concerns exist in data modeling, and the following should be considered when developing an appropriate risk adjustment model.

Sufficient data

When creating a risk adjustment model, there should be enough data available to ensure a stable model. Different statistical rules apply to different types of models. For example, a model with an outcome that is not particularly rare may require more than 30 cases per patient factor in order to consistently return the same model statistics across samples. If the outcome is uncommon, then the number of cases required could be much larger.[7] Other factors may also affect the size needed for a sample, such as a lack of variability among risk factors for a small sample that results in collinearity among risk factors and a corresponding increase in the stability of the parameter estimates. A statistician will be needed to provide guidance to determine the appropriate sample sizes based upon the characteristics of the sample(s) and the requirements of the types of analysis being used.

Model simplicity

Whenever possible, fitting a model with as few variables as possible to explain the most variance possible is preferred. This is often referred to as model simplicity or model parsimony, whereby use of a smaller number of variables accomplishes *about* the same goal as a model with a larger number of

[7] Lezzoni LI. *Risk Adjustment for Measures Health Care Outcomes—Third Edition*. Chicago: Health Administration Press; 2002.

variables is preferred. This principle of parsimony captures the balance between errors of under fitting and over fitting inherent in risk-adjustment model development. For example, developing a model with many predictors can result in model variables that primarily explain incremental variance unique to a data source or available samples (over fitting), and can also result in reduced stability of parameters due to increased collinearity or multicollinearity among a larger number of predictors. In contrast, a model with fewer predictors may reduce the amount of unexplained variance possible for the measure (underfitting).

When evaluating these models, determination of the preferred model may depend upon the availability of other samples to validate findings and detect over fitting, and the degree of multicollinearity among predictors. However, in general, the simpler model may provide a more powerful explanation, since it uses fewer variables to explain *nearly* the same observed variability.[8] In addition, it is likely to reduce the cost of model development by collecting fewer variables, and it may be less likely to show signs of model over fitting. Parsimonious models are often achieved by omitting statistically significant predictors that offer little improvement, and by combining clinically similar conditions to improve performance of the model across time and populations.

Methods to retain/remove risk adjustors

When developing a risk adjustment model, the choice of variables to be included often depends upon estimated parameters in the sample, rather than the true value of the parameter in the population. Consequently, when selecting variables to retain/exclude from a model, the idiosyncrasies of the sample, as well as factors such as the number of candidate variables and correlations among the candidate variables may impact the final risk adjustors retained in a model.[9] Improper model selection or not accounting for the number of or correlation among the candidate variables may lead to risk adjustment models that include suboptimal parameters or overestimated parameters, making them too extreme or inappropriate for application to future datasets. This outcome is sometimes referred to as model overfitting, particularly when the model is more complicated than needed and describes random error instead of an underlying relationship. Given these possibilities, it is advisable to consider steps to adjust for model overfitting, such as selection of model variables based upon bootstrap analysis and assessment of the model in multiple/diverse samples (refer also to Generalizability below). Consultation of clinical expertise, ideally used during candidate variable selection, is also strongly recommended when examining the performance of candidate variables in the risk adjustment models. This expertise may help inform relationships among model parameters, and may help justify decisions to retain or remove variables.

Generalizability

Steps to ensure findings can be generalized to target populations should also be taken when developing the model. Researchers often use two datasets in building risk adjustment models: a

[8]In situations with high visibility or potentially wide-spread fiscal repercussions, CMS has employed some of the most sophisticated models available, such as Hierarchical Generalized Linear Models (Statistical Issues in Assessing Hospital Performance, Commissioned by the Committee of Presidents of Statistical Societies, November 28, 2011).

[9]Harrell, Frank E. *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression.* New York: Springer. 2001.

development (or calibration) dataset and a validation dataset. The development (or calibration) dataset is used to develop the model (or calibrate the coefficients), and the validation dataset is used to determine the extent to which the model can be appropriately applied to other populations. When assessing generalizability to the population from which the development dataset was derived, the two datasets may be collected independently (which can be costly), or one dataset may be split using random selection. Either of these methods allows evaluation of the model's generalizability to the population, and helps avoid any model features that arose from idiosyncrasies in the development sample. Additional validation using samples from different time periods may also be desirable to examine the stability of the model over time. Similarly, depending upon the limitations of the development sample, examination of the model using samples from other diverse populations or data sources may also enhance confidence in the generalizability of the model (e.g., examination of the model's performance in high-volume versus low-volume provider groups, urban versus rural beneficiaries, or administrative versus chart data where both sources exist for the same sample). Across these samples, the model's performance in predicting observed data helps to calibrate model parameters to ensure it will adequately perform when applied to a target population.

Multilevel (hierarchical) data

The potential for observations to be "nested" within larger random groupings (or levels) frequently occurs in healthcare measurement (e.g. patients may be nested under physician groups, who may in turn, be nested under hospitals). The risk adjustment model should account for these multilevel relationships, when present, and risk adjustment development should investigate theoretical and empirical evidence for potential patterns of correlation in this multilevel data. For example, patients in the same IRF may tend to have similar outcomes based upon a variety of factors, and this should be addressed by the risk adjustment model.

Such multilevel relationships are often examined by building models designed to account for relationships between observations within larger groups. Terms for these types of models include multilevel model, hierarchical model, random effects models, random coefficient model, and mixed model. These terms all refer to models that explicitly model the "random" and "fixed" variables at each level of the data. In this terminology, a "fixed" variable is one that is assumed to be measured without error, where the value/characteristic being measures is the same across samples (e.g. male versus female, non-profit versus for-profit facility) and studies. In contrast, "random" variables are assumed to be values drawn from a larger population of values (e.g. a sample of Inpatient Rehabilitation Facilities), where the value of the random variable represents a random sample of all possible values of that variable.

Traditional statistical methods (such as linear regression and logistic regression) require observations (e.g., patients) in the same grouping to be independent. When observations covary based upon the organization of larger groupings, these methods fail to account for the hierarchical structure, and assumptions of independence among the observation are violated. This may ultimately lead to underestimated standard errors, and lead to incorrect inferences. Attempts to compensate for this problem by treating the grouping units as fixed variables within a traditional regression framework are

generally undesirable, as the grouping units must be treated as a fixed variable, which does not allow for generalization to any other groupings beyond those grouping units in the sample.

Multilevel models overcome these issues by explicitly modeling the grouping structure, and by assuming that the groups reflect random variables (usually with a normal distribution) sampled from a larger population. They take into account variation at different grouping levels, and allow modeling of hypothesized factors at these different levels (e.g., a multilevel model may allow modeling patient-level risk factors as well as facility level factors). If the measure developer has reason to suspect hierarchical structure in the measurement data, these models should be examined. They can be applied within common frameworks used for risk adjustment (e.g., ordinary least squares regression for continuous outcomes, logistic regression for binary outcomes), as well as less common longitudinal frameworks such as growth (change) modeling.

Developments in statistics are enabling researchers to improve both the accuracy and the precision of nested modes using computer-intensive programs recently available. These models include estimation of clustering effects independent of the main effects of the model to better evaluate the outcome of interest. For example, the use of precision-weighted empirical Bayesian estimation has been shown to produce more accurately generalizable coefficients across populations than methods that rely on the normal curve for estimation (e.g., linear regression). Hierarchical factor analysis and structural equation modeling have also been used. Recently, CMS has moved towards using one of these models (i.e., Hierarchical Generalized Linear Model) for monitoring and reporting hospital readmissions.[10]

Step 6: Assess the model

This step is required for a risk adjustment model developed *de novo,* and it is also required when using an "off the shelf" adjustment model because an existing risk adjustment model may perform differently in the new measure context. When multiple data sources are available (e.g. administrative and chart-based data), it is strongly recommended that model performance is assessed for each data source to allow judgment regarding the adequacy and comparability of the model across the data sources.

Any model developed should be assessed to ensure that it does not violate underlying model assumptions (e.g., independence of observations or assumptions about underlying distributions) beyond the robustness established in the literature for those assumptions. Models must also be assessed to determine the predictive ability, discriminant ability, and overall fit of the model, and justification of the types of models used must be provided to the Contracting Officer Representative/Government Task Leader, as well as documented in the Risk Adjustment Methodology report. Some examples of common statistics used in assessing risk adjustment models include the R^2 statistic, receiver operating characteristic (ROC), and Hosmer-Lemeshow test. However, several other statistical techniques exist that allow developers to assess different aspects of model fit for different

[10]CMS Statistical White Paper-*Issues in Assessing Hospital Performance*, November 28, 2011. Available at http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf. Accessed August 28, 2012.

subpopulations as well as for the overall population. Use of an experienced statistician is critical to ensure the most appropriate methods are selected during model development and testing.

R^2 statistic

A comparison of the R^2 statistic with and without selected risk adjustment is frequently used to assess the degree to which specific risk-adjusted models predict, explain, or reduce variation in outcomes unrelated to an outcome of interest. The statistic can also be used to assess the predictive power of risk-adjusted models, overall. In that case, values for R^2 describe how well the model predicts the outcome based on the values of the included risk factors.

The R^2 value for a model can vary, and no firm standard exists for what is the best value. One may depend on what seems "reasonable" in an R^2 value based upon past experience or previously developed models. In general, the larger the R^2 value, the better the model. However, clinical expertise may also be needed to help assess whether remaining variation is primarily related to differences in the quality being measured.

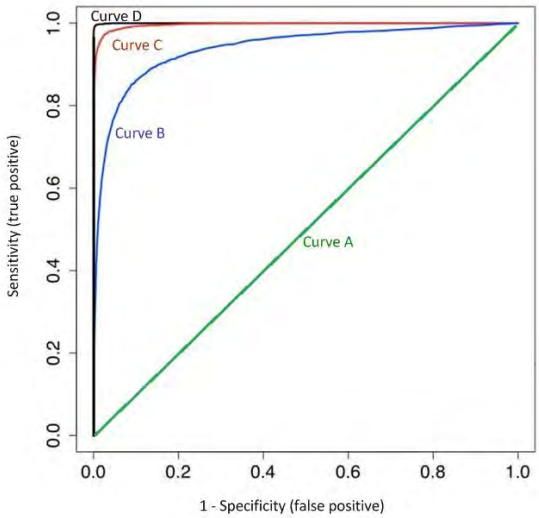
ROC curve, AUC, and C-statistic

A ROC curve is often used to assess models that predict a binary outcome (e.g., a logistic regression model), where responses are classified into two categories. The ROC curve can be plotted as the proportion of target outcomes correctly predicted (i.e., a true positive) against the proportion of outcomes incorrectly predicted (i.e., a false positive). The curve depicts the tradeoff between the model's sensitivity and specificity.[11] An example is shown if Figure 11-1. Curves approaching the 45-degree diagonal of the graph represent less desirable models (see curve A) when compared to curves falling to the left of this diagonal that indicate higher overall accuracy of the model (see curves B and C). A test with nearly perfect discrimination will show a ROC curve that passes through the upper-left corner of the graph, where sensitivity equals 1 and 1 minus specificity equals zero (see curve D).

The power of a model to correctly classify outcomes into two categories (i.e., discriminate) is often quantified by the area under the ROC curve (AUC). The AUC, sometimes referred to as the c-statistic, is a value that varies from 0.5 (discriminating power not better than chance) to 1.0 (perfect discriminating power). It can be interpreted as the percent of all possible pairs of observed outcomes in which the model assigns a higher probability to a correctly classified observation than to an incorrect observation. Most statistical software packages compute the probability of observing the model AUC found in the sample when the population AUC equals 0.5 (the null hypothesis). Both non-parametric and parametric methods exist for calculating the AUC, and this varies by statistical software.

[11] Sensitivity is the ability to correctly predict or find a condition of interest, and specificity is the ability to correctly predict or find the lack of a condition of interest.

Figure 11-1 Example of ROC curves



The Hosmer-Lemeshow test

While the AUC/c-statistic values provide a method to assess a model's discrimination, the quality of a model can also be assessed by how closely the predicted probabilities of the model agree with the actual outcome (i.e., whether predicted probabilities are too high or too low relative to true population values). This is sometimes referred to as *calibration* of a model. It is often assessed using the Hosmer-Lemeshow test of goodness-of-fit, which assesses the extent to which the observed values/occurrences match expected event rates in subgroups of the model population. The Hosmer-Lemeshow test identifies subgroups of ordered observations based upon the predicted model values or other factors external to the model associated with the outcome risk. The subgroups can be formed for any reasonable grouping, but often deciles or quintiles are used. Generally, a model is considered well calibrated when the expected and observed values agree for any reasonable grouping of the observations. Yet, high risk and low frequency situation pose special problems for these types of comparison methodologies that should be addressed by an experienced statistician.

A statistician with experience in such methodology can determine the adequacy of any model. It is expected that the Measure Contractor team will employ the services of a statistician to accurately assess the appropriateness of a risk-adjusted model. Determining the "best" risk-adjusted model may involve multiple statistical tests that are more complex than what is cited here. For example, a risk adjustment model may discriminate very well based upon the c-statistic, but still be calibrated poorly. Such a model may predict well at low ranges of outcome risk for patients with a certain set of characteristics (e.g., the model produces an outcome risk of 0.2 when roughly 20 percent of the patients with these characteristics exhibit the outcome in population), but predict poorly at higher ranges of risk (e.g., the model produces an outcome risk of 0.9 for patients with a different pattern of characteristics when only 55 percent of patients with these characteristics show the outcome in population). In this case, one or more goodness-of-fit indices may need to be consulted to identify a superior model, and careful analysis of different subgroups in the sample may also be needed to

further refine the model. Additional steps to correct for bias in estimators, improving confidence intervals, and assessing any violation of model assumptions may also be required. Moreover, the differences across groups for measures that have not been risk adjusted may be clinically inconsequential when compared to risk-adjusted outcomes, and clinical experts in the subject matter at hand for the Measure Contractor team are also expected to be consulted (or employed) to provide an assessment of both the risk adjustors and utility of the outcomes.

Step 7: Document the model

A Risk Adjustment Methodology report is considered a required deliverable, and it is expected upon the conclusion of a measure development project. This report ensures that relevant information about the development and limitations of the risk adjustment model are available for review by consumers, purchasers, and providers. It also allows these parties to access information about the factors incorporated into the model, the method of model development, and the significance of the factors used in the model. Typically the report will contain:

- Background.
- Identification or review of the need for risk adjustment of the measures.
- A description of the sample(s) used to develop the model, including criteria used to select the sample and/or number of sites/groups, if applicable.
- A description of the methodologies and steps used in the development of the model, or a description of the selection of an "off the shelf" model.
- A listing of all variables considered and retained for the model, the contribution of each retained variable to the model's explanatory power, and a description of how each variable was collected (e.g., data source, time frames for collection).
- Description of the model's performance, including any statistical techniques used to evaluate performance, and a summary of model discrimination and calibration in one/more samples.
- Important limitations are delineated, such as the probable frequency and impact from misclassification when the model is used. For example, classifying a high outcome provider as a low one or the reverse.[12]
- Enough summary information about the comparison between unadjusted and adjusted outcomes to allow an evaluation of the clinical significance of the model's impact.
- A section discussing a recalibration schedule for the model to accommodate changes in medicine and in populations. Such schedules are normally first assigned based on the experience of clinicians and the literature's results and later updated as needed.

All measure specifications, including the risk adjustment methodology, must be fully disclosed. The risk adjustment method, data elements, and algorithm are to be fully described in the Risk Adjustment portion of the Measure Information form. Attachments or links to Web sites should be provided for coefficients, equations, codes with descriptors, and definitions and/or specific data collection items/responses used in the risk adjustment. Documentation should comply with NQF's open source

[12]Austin PC, Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals, *BMC Med Res Methodol,* 2008;8:30.

requirements and all applicable programming code should be included. If calculation requires database-dependent coefficients that change frequently, the existence of such coefficients and the general frequency that they change should be disclosed, but the precise numerical values assigned need not be disclosed because they vary over time.

12. Public Comment

12.1 Introduction

The public comment period ensures that the Centers for Medicare & Medicaid Services (CMS) measures continue to be of the highest caliber possible by using a transparent process with balanced input from relevant stakeholders.

The public comment period provides an opportunity for the widest array of interested parties to provide input on the measures under development and to provide critical suggestions not previously considered by the measure contractor or technical expert panel (TEP).

The federal rule-making process includes a public comment period and is another method in which feedback is often obtained for CMS measures. The rulemaking process is primarily used when measures are related to reimbursement programs.

12.2 Deliverables

- Call for Public Comment Web posting form
- List of stakeholders for notification
- Measure Information forms (MIF) and Measure Justifications (for candidate measures)
- Public Comment Summary report, including verbatim comments

12.3 Procedure

The following steps are essential to soliciting public comment. Deviation from the following procedure requires the Contracting Officer Representative/Government Task Leader's (COR/GTL's) approval.

The call for public comment involves several postings to the dedicated CMS Measures Management System (MMS) web site at: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/CallforPublicComment.html. The measure contractors will work with the Measures Manager to post the call. Web site postings involve two CMS divisions, and the process to post each set of materials will take approximately five working days.

Step 1: Write the call for public comment

Prepare the Call for Public Comment Web posting document. Measure contractors may use this document as a means of soliciting public comment on CMS measures. This document includes general information regarding the purpose of the call for comments and instructions on how to submit comments. A template of the Call for Public Comments Web posting is located at the end of this section.

When organizing a Call for Public Comment, arrange for an e-mail address to receive the comments.

Step 2: Notify relevant stakeholder organizations

Submit a list of relevant stakeholder organizations for notification about the public comment period to the COR/GTL for review and input prior to posting the call. After approval by CMS, notify the stakeholder organizations before the Web site posting goes live. Relevant stakeholder groups may include, but are not limited to:

- Quality alliances (AQA, PQA, and others).
- Medical societies.
- Scientific organizations related to the measure topic.
- Measure developers (AMA-PCPI, NCQA, The Joint Commission, RAND Corporation, etc.).
- Other CMS measure contractors.
- Consumer organizations.
- Quality improvement organizations (QIOs)/ESRD networks.
- Purchaser groups.
- Impacted provider groups/professional organizations.
- Individuals with quality measurement expertise.
- Individuals with health disparities measurement expertise.

Notification methods may include, but are not limited to:

- Press releases.
- Notice to the CMS Communications Coordinator with CMS Web and New Media Group as part of The Office of Communications.
- Notice via e-mail to the stakeholder mailing list.
- Social media (e.g., Twitter, Facebook, YouTube, LinkedIn, etc.) Contact your COR/GTL for the process.

Step 3: Post the measures following COR/GTL's approval

After obtaining the COR/GTL's approval, work with the Measures Manager to post the Measure Information and Measure Justification forms on the dedicated Web site. https://www.cms.gov/MMS/17_CallforPublicComment.asp. The steps in the posting process are described later in this section. When submitting the forms, prominently mark the MIFs and Measure Justification forms as "draft."

As a general rule, the call should be posted on the Web site for at least two weeks to allow sufficient time for the public to provide comments. The COR/GTL will make the final decision as to how long the call should be posted.

In the event the dedicated Web site is temporarily unavailable, consult with the COR/GTL and/or the Measures Manager to investigate other electronic communication methods.

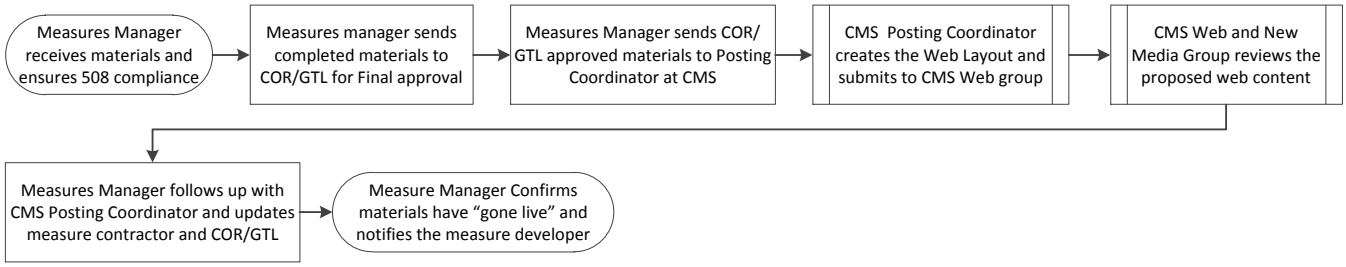
The information to be posted should include:

- The specific objectives of the measure development contract.
- The processes used to develop the measures, for example:

- Identifying important quality goals related to Medicare services.
- Conducting literature reviews and grading evidence.
- Defining and developing specifications for each quality measure.
- Obtaining evaluation of proposed measures by technical expert panels (as directed by the COR/GTL, the TEP Summary report may be posted)
- Measure testing.
- Posting for public comment.
- Refining measures as needed.

- The Measure Information Form and Measure Justifications.
- A list of the TEP members, including any potential conflicts of interest disclosed by the members.
- Information about the measure contractor and subcontractors developing this measure set.

Figure 12-1 The Posting Process



The posting process:

1. Upon receipt of the posting materials, the Measures Manager reviews to confirm 508 compliance.
2. Measures Manager will request final approval from the COR/GTL to send to the CMS Web site Posting Coordinator for posting.
3. Measures Manager sends the materials to the CMS Web site Posting Coordinator for loading.
4. CMS Web site Posting Coordinator creates the updated Web page layout and submits it to the CMS Web group for posting.
5. CMS Web and New Media Group, as part of The Office of Communications, is responsible for the entire CMS Web site. They review the proposed Web content to ensure it meets all CMS Web site requirements. It is then moved to the production environment where the page "goes live."
6. Measures Manager receives confirmation from the CMS Web site Posting Coordinator when the approved materials have been moved into the production environment.
7. Measures Manager confirms that the materials are available on the site and notifies you and the COR/GTL.

If a relatively quick turnaround time is required for timely posting, it is best for you to ask the COR/GTL to monitor the process. The posting process can take up to 5 business days and should be factored into their timeline. Note: materials sent at the end of a business day may not be reviewed until the next business day.

Step 4: Collect information

Commenters will submit their comments via email or other Web-based tool (e.g., Survey Monkey) as directed on the CMS MMS Web site. The public is encouraged to submit general comments on the entire measure set or comments specific to certain measures.

The Paperwork Reduction Act (PRA) mandates that all federal government agencies must obtain approval from the Office of Management and Budget (OMB) before collection of information that will impose a burden on the general public. Measure contractors should be familiar with the PRA before implementing any process that involves the collection of new data. Contractors should consult with their GTL/COR regarding the PRA to confirm if OMB approval is required before requesting most types of information from the public. The full Act is available online at http://www.archives.gov/federal-register/laws/paperwork-reduction/.

HHS also has an additional Web site with frequently asked questions and answers http://www.hhs.gov/ocio/policy/collection/infocollectfaq.html. The following Question and Answer appears on the HHS site:

> *Q. Do you need PRA clearance if you just ask people for comments on a document or public comments through the Federal Register?*
>
> *A. Not unless, respondents are asked to respond to specific questions in their comments. If the comment is very general, the PRA doesn't apply. Please note that general public comments can provide limited data and will work well if the program just wants to identify a perceived issue or concern. However, since the responses are limited to what the respondent wants to share with the requestor, useful unbiased data for use at the policy making or research level cannot be obtained from public comments alone.*

Step 5: Summarize comments and produce report

At the end of the public comment period, prepare a preliminary Public Comment Summary Report including verbatim comments to summarize and analyze the public comments received. A template for this report is located at the end of this section. Preliminary recommendations may be stated in the report pending discussion with the TEP. This report should be submitted to the COR/GTL and the TEP within two weeks following the end of the public comment period.

The report should include a:

- Summary of general comments posted and any other information that could apply to the set of measures and recommended action.
- Summary of the comments for each measure and any preliminary recommendations for TEP consideration.
- Listing of the verbatim public comments. If the submitter includes personal health information in relation to the measure, the measure contractor should obscure or remove the sensitive portions prior to posting or releasing the report.

Time permitting, and upon approval from the COR/GTL, work with the Measures Manager to post the preliminary Public Comment Summary Report including verbatim comments on the CMS Web site.

Step 6: Send comments to TEP for consideration

Reconvene the TEP to discuss the submitted comments and preliminary recommended actions. After deliberations, the TEP may make recommendations to the measure contractor concerning changes to the measures as a result of the public comments. This may be done by e-mail, teleconference, or an in-person meeting.

Step 7: Finalize the Public Comment Report including verbatim comments

Document the TEP discussion and the recommended actions. The finalized report should include:

- Recommendations and actions taken in response to the comments received.
  - e.g., candidate measures recommended to be eliminated from further consideration.
- Updated or revised measure specifications with notations about changes made.

Submit to the COR/GTL within one week after the TEP meeting to review the comments.

Step 8: Arrange for Public Comment Summary report to be posted on the Web site

After obtaining the COR/GTL's approval, work with the Measures Manager to post the Public Comment Summary Report including verbatim comments within three weeks (or as directed by the COR/GTL) after the public comment period closes. The posting process is the same as described in Figure 12-1 The Posting Process. This posting will remain on the Web site for a minimum of 21 days. The process for posting the summary report is described earlier in the section. After the 21-day period, the posting may be removed from the Web site.

Call for Public Comment West Posting Template

Project overview:

Centers for Medicare & Medicaid Services (CMS) has contracted with *<contractor name>* to develop *<measure set name or description>*. The purpose of the project is to develop quality measures that can be used to provide quality care to Medicare beneficiaries. The public comment period provides an opportunity for the widest array of interested parties to provide input on the measures under development and can provide critical suggestions not previously considered by the measure contractor or its technical expert panel (TEP).

Specific project objectives: *<list contract objectives>*

The development process includes:

- <Identifying important quality goals related to a topic/condition or setting of focus>
- Conducting literature reviews and grading evidence >
- Defining and developing specifications for each quality>y measure
- Obtaining evaluation of proposed measures by technical> expert panels
- Posting for public comment >
- Testing measures for reliability, validity, and feasibility

To provide public comments, note the following:

- The public is encouraged to submit general comments on the entire measure set or comments specific to certain measures.
- At the end of the public comment period, all public comments will be posted on the Web site.
- Do not include personal health information in your comments.

Instructions for Providing Comments:

- If you are providing comments on behalf of an organization, include the organization's name and your contact information.
- If you are commenting as an individual, submit identifying or contact information.
- Please indicate which measures you are providing comments on. You may submit general comments on the entire set of measures or you may provide comments specific to individual measures.
- Email your comments to: *<insert email address>*. Comments are due by close of business *<insert date>*.
- *(Measure Contractor shall provide the following for <u>each</u> measure):*
- *<Measure A-Measure Name>.*
- *<Measure B-Measure Name>.*
- *<Measure C-Measure Name>* etc.

Template for Public Comment Summary Report including verbatim comments

<Project Name>

<Date of Report>

<Contractor Name>

Introduction

- Dates of public comment period
- Web site used.
- Methods used to notify stakeholders and general public of comment period.
- Volume of responses received.

Stakeholder Comments—General

- Summary of general comments.
- Proposed action(s).

Measure-Specific Comment Summaries ***(Complete this section for each measure)***

- Measure name.
- Summary of comments.
- Proposed action(s).

Preliminary Recommendations (this section can be deleted after the TEP discussion)

Overall Analysis of the Comments and Recommendations to CMS (include a summary of the TEP discussion and changes to the list of candidate measures)

Verbatim Comments

Date Posted	Measure Set or Measure	Text of Comments	Name, Credentials, and Organization of Commenter	E-Mail Address	Type of Organization	Recommendations/Actions Taken

13. Measure Testing

13.1 Introduction

Measure testing enables a measure developer to assess the suitability of the measure's technical specifications, and acquire empirical evidence to help assess the strengths and weaknesses of a measure with respect to the evaluation criteria. This information can be used in conjunction with expert judgment to evaluate a measure.

Properly conducting measure testing and analysis is critical to approval of a measure by The Centers for Medicare & Medicaid Services (CMS) and endorsement by the National Quality Forum (NQF). This section describes the types of testing that may be conducted during measure development (alpha and beta testing), the procedure for planning and testing under the direction of the CMS Contracting Officer Representative/Government Task Leader (COR/GTL), and key considerations when analyzing and documenting results of testing and analysis.

The information in this section is not meant to be prescriptive or exhaustive. Other approaches to testing that employ appropriate methods and rationale may be used. Measure developers should always select testing that is appropriate for the measure being developed, and always provide empirical evidence for importance to measure and report, feasibility, scientific acceptability, and usability and use.

13.2 Deliverables

- Measure Testing Plan
- Measure Testing Summary report
- Updated Measure Information Form (MIF)
- Updated Measure Justification form
- Updated Measure Evaluation report

13.3 Alpha and Beta Testing

A measure should be tested one or more times during the development process. Testing provides an opportunity to refine the draft specifications before they are finalized; augment or re-evaluate earlier judgments about the measure's importance; and assess the feasibility, usability, and scientific acceptability of the measure. Initial testing during development (sometimes referred to as pilot testing) is generally conducted within the framework of alpha and beta tests. Attributes of each type of test are shown below, and these may be used as considerations when developing a work plan for alpha or beta tests.

Alpha testing

Alpha tests (also called formative tests) are of limited scope. They usually occur during the formative stage of measure development before detailed specifications are fully developed. The primary purpose of alpha testing is to refine or confirm the utility of envisioned technical specifications, and to conduct

an initial assessment of measure feasibility. The types of testing done in an alpha test vary widely, and often depend upon the measure's data source or novelty of the specifications for the measure. Measures that use specifications similar to existing measures may require very little alpha testing. In contrast, measures that address areas for which specifications have never been developed may require multiple iterations of an alpha test.

Beta testing

Beta testing (also called field testing) generally occurs after the initial technical specifications have been developed, and is usually larger in scope than alpha testing. In addition to gathering further information about feasibility, beta tests serve as the primary means to assess scientific acceptability and usability of a measure. They can also be used to evaluate the measure's suitability for risk adjustment/stratification, and help expand prior evaluations of the measure's importance and feasibility. Careful planning and execution of beta testing facilitates reporting and documenting measure properties with respect to criteria used by CMS and NQF.

Table 13-1 compares key features of alpha and beta testing.

Table 13-1 Features of Alpha and Beta Testing

	Alpha Testing	Beta Testing
Timing	◆ Usually carried out prior to the completion of technical specifications. ◆ May be carried out multiple times in quick succession.	◆ After the measure contractor detailed and precise technical specifications developed
Scale	◆ Typically smaller scale ◆ Only enough records to ensure data set contains all elements needed for the measure ◆ Only enough records to identify common occurrences or variation in the data	◆ Strives to achieve representative sample sizes. ◆ Requires appropriate sample selection protocols. ◆ May require evaluation of multiple sites in a variety of settings depending upon the data source (e.g., administrative, medical chart).
Sampling	◆ Convenience sampling	◆ Sufficient to allow adequate testing of the measure's scientific acceptability ◆ Representative of the target population ◆ Representative of the people, places, times, events, and conditions important to the measure. ◆ If based on administrative data use the entire eligible population.
Specification Refinement	◆ Permits the early detection of problems in the technical specifications (e.g., identification of additional inclusion and exclusion criteria).	◆ Used to assess or revise the complexity of computations required to calculate the measure.

	Alpha Testing	Beta Testing
Importance	◆ Designed to look at the volume, frequency, or costs related to a measure topic (i.e., cost of treating the condition, costs related to procedures measured, etc.) ◆ Establish on a preliminary basis that the measure can identify low levels of care quality. ◆ Provides support for further development of the measure.	◆ Allows for enhanced evaluation of a measures importance including evaluation of performance thresholds and outcome variation. ◆ Evaluate opportunities for improvement in the population, which aids in evaluation of the measure's importance (e.g., obtaining evidence of substantial variability among comparison groups; obtaining evidence that the measure is not "topped out" where most groups achieve similarly high performance levels approaching the measure's maximum possible value).
Scientific Acceptability	◆ Limited in scope if conducted during the formative stage. Usually occurs later in development.	◆ Assesses measure reliability and validity. ◆ Report results of analysis of exclusions (if any used). ◆ Test results of risk adjustment model, quantifying relationships between and among factors.
Feasibility	◆ Identifies barriers to implementation. ◆ Provides initial information about the feasibility of collecting the required data and calculating the measures using the technical specifications. ◆ Offers an initial estimate of the costs or burden of data collection and analysis.	◆ Provides enhanced information regarding feasibility including greater determination of the barriers to implementation and costs associated with measurement. ◆ Evaluates the feasibility of stratification factors based upon occurrences of target events in the sample, or be used to inform risk adjustment decisions. ◆ Identify unintended consequences, including susceptibility to inaccuracies and errors. ◆ Report strategies to ameliorate unintended consequences.
Usability	◆ No formal analytic testing at this stage. The TEP may be used to assess the potential usability of the measure.	◆ May consist of focus groups or similar means of assessing usefulness of the measure by consumers. This type of testing is often not in the scope of measure development contracts. ◆ The TEP may also be used to assess the potential usability.

Sampling

The need for careful sampling often varies depending upon the type of test (alpha or beta) and the type of measure. For example, measures that rely on administrative data sources (e.g., claims) can be tested by examining data from the entire eligible population with limited impact on external resources. However, to test some measures it is necessary to collect information from service providers and/or beneficiaries directly, which can become burdensome. As noted above, alpha testing frequently uses a sample of convenience while beta testing may involve measurement of a target population which requires careful construction of samples to support adequate testing of the measure's scientific acceptability. The analytic unit of the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy. In general, samples used for reliability and validity testing should:

- Represent the full variety of entities whose performance will be measured (e.g., large and small hospitals). This is especially critical if the measured entities volunteer to participate, which limits generalizability to the full population.

- Include adequate numbers of observations to support reliability and validity analyses using the planned statistical methods.
- When possible, observations should be randomly selected.

However, when determining the appropriate sample size during testing it is necessary to evaluate the burden placed on providers and/or beneficiaries to collect the information. The Paperwork Reduction Act (PRA) mandates that all federal government agencies obtain approval from the Office of Management and Budget (OMB) before collection of information that will impose a burden on the general public. Measure contractors should be familiar with the PRA before implementing any process that involves the collection of new data. As such, contractors must maintain an appropriate balance between the collection of information to support the reliability and validity of its measures and the burden created by its collection. Once a measure has been implemented and data has been collected for reporting, more rigorous sampling and measure testing can be conducted during the maintenance phase (refer to Volume 2).

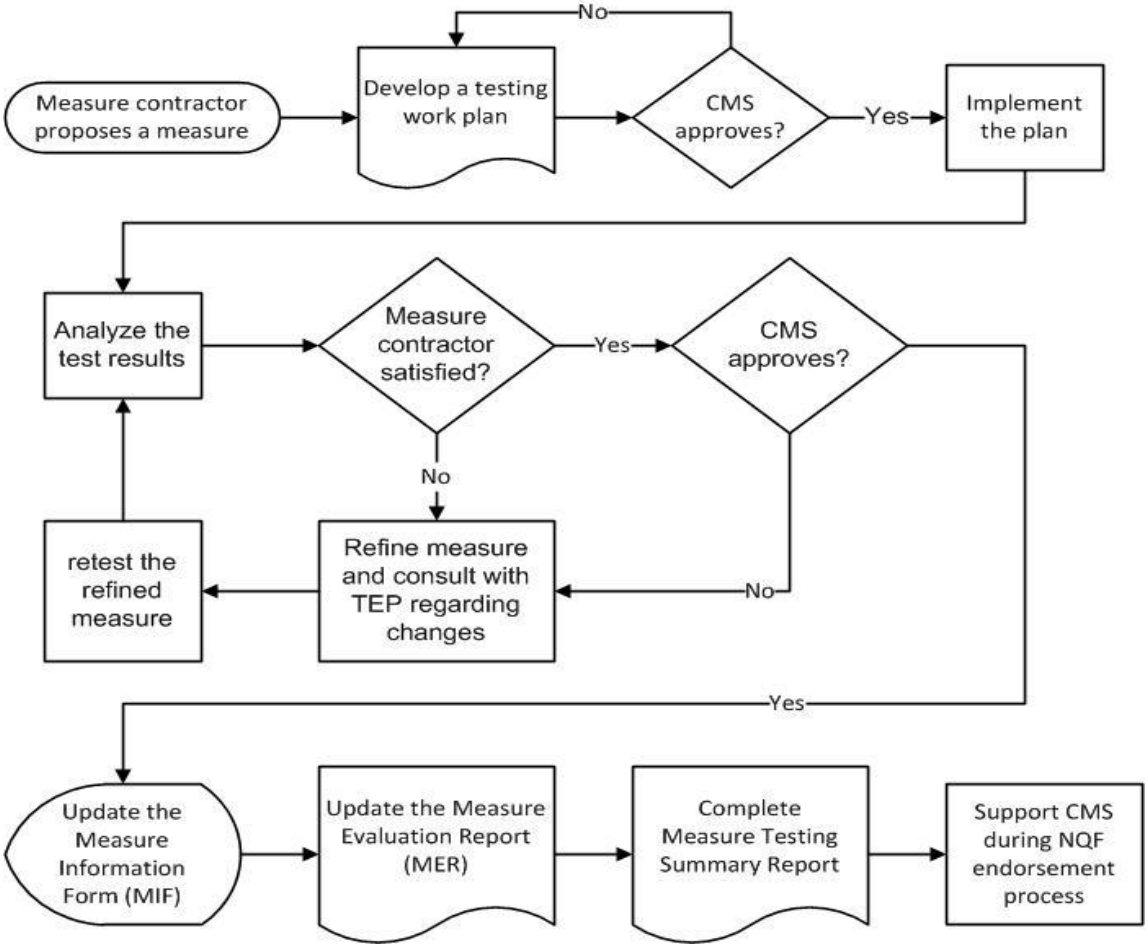
Contractors should consult with their COR/GTL regarding the PRA to confirm if OMB approval is required before requesting most types of information from the public. The full Act is available online at http://www. archives. gov/federal-register/laws/paperwork-reduction/. HHS also has an additional Web site with frequently asked questions and answers http://www. hhs. gov/ocio/policy/collection/infocollectfaq. html.

13.4 Procedure

When developing a measure (or set of measures) for CMS, a measure developer is required to submit specific reports, and is encouraged to follow the steps listed below. Steps 1–6 address planning and implementation of the testing, and these are identical for alpha or beta testing while Steps 7–11 address reporting and follow-up after the conclusion of testing. While this always applies to the completion of reports following beta testing, measure developers should discuss the need for reporting upon more formative alpha testing with the COR/GTL, especially if the alpha testing is intended to precede beta testing under the same measure development contract.

An overview of the procedure is shown in Figure 13-1.

Figure 13-1 Overview of Measure Testing



Step 1: Develop the testing work plan

Measure testing can be conducted for a single measure or a set of measures. If the testing targets a set of measures, construct a work plan that describes the full measure set. The work plan for alpha testing is usually prepared early in the measures development process; therefore, the exact number of measures to be tested may not be known, and many of the work plan areas listed below may not be appropriate. In contrast, the work plan for a beta test should be prepared after the measure specifications have been developed, and it should include sufficient information to help the COR/GTL understand how the sampling and planned analyses ensure that the measures meet scientific acceptability, usability, and feasibility criteria required for approval by CMS and endorsement by NQF.

The testing plan usually contains the following:

- Name(s) of measure(s).
- Type of testing (alpha or beta).
- Study objective(s).
- The timeline for the testing and report completion.

- Data collection methodology:
- Description of test population; include number and distribution of test sites/data sets.
  - Description of the data elements that will be collected.
  - Sampling methods to be used (if applicable).
  - Description of strategy to recruit providers/obtain test data sets (if multiple sites or data sets are used).
- Analysis methods planned, and a description of test statistics that will be used to support assessment. This will be less extensive for an alpha test. For a beta test, methods and analysis should address the following four evaluation criteria:
  - Importance—including analysis of opportunities for improvement such as variability in comparison groups or disparities in health care related to race, ethnicity, age or other classifications.
  - Scientific acceptability—including analysis of reliability, validity, and exclusion appropriateness.
  - Feasibility—including evaluation of reported costs or perceived burden, frequency of missing data, and description of data availability.
  - Usability—including planned analyses to demonstrate that the measure is meaningful and useful to the target audience. This may be accomplished by obtaining review of measure results (e.g., means and detectable differences, dispersion of comparison groups, etc.) to the technical expert panel (TEP) for review. More formal testing, if requested by CMS, may require assessment via structured surveys or focus groups to evaluate the usability of the measure (e.g., clinical impact of detectable differences, evaluation of the variability among groups, etc.).
- Description and forms documenting patient confidentiality, and description of institutional review board (IRB) compliance approval or steps to obtain data use agreements (if necessary).
- Methods to comply with the Paperwork Reduction Act.
- Training and qualification of staff. For example, identifying those who will do the following:
  - Manage the project (and their qualifications).
  - Conduct the testing (and their qualifications).
  - Conduct or oversee data abstraction.
  - Conduct or oversee data processing.
  - Conduct or oversee data analysis.

Step 2: Submit the work plan to the COR/GTL for approval

Submit the work plan to the COR/GTL with any necessary supporting documents.

Step 3: Implement the approved work plan

Following COR/GTL review and approval, implement the work plan.

Step 4: Analyze the testing results

Once all data are gathered from the test sites, the developer conducts a series of analyses in order to characterize the feasibility, integrity and face validity of the measures being tested. The findings of all testing analyses will be presented in a summary final report described in Step 10 and discussed with the COR/GTL.

Step 5: Refine the measure

Based upon the analysis results, modification may be required for the measure specifications, data collection instructions, and calculation of measure results. For example:

- Following alpha testing, measure re-specification or efforts to overcome implementation barriers are often undertaken.
- Following beta testing, changes in the definition of the population or adjustments to the comparison group definition may occur.
- If changes to the measure are made, consultation with the TEP is recommended prior to retesting the measure.

Step 6: Retest the refined measure

Continue to refine and retest measures as deemed necessary by the measure contractor and the COR/GTL.

Step 7: Review findings with CMS

Communicate findings to CMS for review. Findings should be communicated based upon the preference of the COR/GTL. Based upon COR/GTL instructions, complete additional reporting forms in Steps 8–10.

Step 8: Update the Measure Information Form

Update the MIF with revised specifications, and update the Measure Justification form with new information obtained during testing, including: additional information about importance such as variability in comparison groups and opportunities for improvement; reliability, validity, and exclusion results; risk adjustment or stratification decisions; usability findings; and feasibility findings.

Step 9: Update the Measure Evaluation report

For each measure, based on the results from beta testing, prepare a Measure Evaluation report to summarize how the measure meets each of the measure evaluation criteria and subcriteria. It is important to evaluate the measure in an as objective manner as possible in order to anticipate any issues when the measure is submitted to NQF for endorsement. The measure evaluation report is where the contractor can communicate any anticipated risks associated with endorsement and present plans to strengthen any weaknesses identified. It is important for CMS to have an understanding of what it would take (pros/cons, costs/benefits) for increasing the rating and the risks if not undertaken.

The Measure Evaluation report can be modified as appropriate. It can be included as part of the Measure Testing Summary report.

Step 10: Submit Measure Testing Summary report

For each measure or set of measures, complete required summary reports and submit to the COR/GTL. Following the analysis of information acquired during testing, the measure contractor must summarize the measure testing findings. With respect to the NQF measure evaluation criteria—Importance to Measure and Report, Scientific Acceptability of Measure Properties, Usability, and Feasibility—the goal of these summaries is to document evidence sufficient to achieve approval by CMS and endorsement by NQF.

When reporting measure testing results, assessment upon each of the four measurement criteria is a matter of degree. For example, not all revisions will require extensive reassessment for all testing criteria, and not all previously endorsed measures will be strong—or equally strong—among each set of criteria. This is often a matter of judgment and expertise. Given the difficulty of assessment, measure contractors are expected to contract or employ experienced statisticians and methodologists to provide expert judgment when reporting measure reliability and validity, and also summarize expert findings/consensus with respect to measure:

- Importance
- Acceptability
- Usability
- Feasibility

The following are recommendations for the content of the Measure Testing Summary report. However, these recommendations are not intended to be exhaustive, and not all recommendations will apply to each measure depending upon the type of testing and the characteristics of the measure.

The summary of testing may include the following information:

- Name of measure, measure set
- An executive summary of the tests and resulting recommendations
- Type of testing conducted (alpha or beta), and an overview of the testing scope
- Description of any deviation from the work plan along with rationale for deviation
- Data collection and management method(s):
  - Description of test population(s) and description of test sites (if applicable)
  - Description of test data elements including type and source
  - Data source description (and export/translation processes, if applicable)
  - Sampling methodology (if applicable)
  - Descriptions of exclusions (if applicable)
  - Medical record review process (if applicable) including abstractor/reviewer qualifications and training, and process for adjudication of discrepancies between abstractors/reviewer
- Detailed description of measure specifications and measure score calculations
- Description of the analysis conducted, including:

- Qualifications of analysts performing tests.
- Summary statistics (e.g. means, medians, denominators, numerators, descriptive statistics for exclusions, etc.).
- *Importance*—Specific analyses demonstrating importance such as suboptimal performance for a large proportion of comparison groups, and analysis of differences between comparison groups.
- *Scientific Acceptability*
  - *Reliability*—Description of reliability statistics and assessment of adequacy in terms of norms for the tests, and the rationale for analysis approach.
  - *Validity*—Specific analyses and findings related to any changes observed relative to analyses reported during the prior assessment/endorsement process, or changes observed based upon revisions to the measure. These may include assessment of adequacy in terms of norms for the tests conducted, panel consensus findings, and rationale for analysis approach.
  - *Exclusions/Exceptions*—Discussion of the rationale (if different from the original measure specifications), which may include listing citations justifying exclusions; documentation of TEP qualitative or quantitative data review; changes from prior assessment findings such as summary statistics and analyses, which may include changes in frequency and variability statistics; and sensitivity analyses.
  - Analysis of the need for risk adjustment and stratification (refer to the Risk Adjustment section).
- *Usability*—If affected by changes following implementation, or the measure has been materially changed, a summary of findings related to measure interpretability and methods used to provide a qualitative and quantitative usability assessment is recommended (e.g., TEP review of measure results; or, in rare situations, use of a CMS-requested focus group or survey).
- *Feasibility*—Discussion of feasibility challenges and adjustments that were made to facilitate obtaining measure results, and description of estimated costs or burden of data collection.

- Any recommended changes to the measure specifications and an assessment as to whether further testing is needed.
- A detailed discussion of testing results compared to NQF requirements, including whether or not the NQF requirements are believed to have been sufficiently met, or if additional testing is required.
- Limitations of the alpha or beta testing
  - e.g., sample limited to two sites or three electronic health record (EHR) applications; sample used registry data from one state, and registry data is known to vary across states; testing was formative alpha test only, and was not intended to address validity and reliability
- Specific to eMeasures:

- The number of test sites reporting each measure data element and each measure overall as feasible to implement; or feasible with workflow changes; or not feasible to implement at this time and relevant comments
- The number of test sites using a specific coding system to record the specific data element (i.e., SNOMED, LOINC, etc.).
- The number of test sites using a specific data capture type for each data element (i.e., discrete/non-discrete, numeric, Boolean, etc.).
- Percentage of feasible elements for each measure per test site (reported as range, the high and low sites).
- Percentage of test sites reporting the measure retains an acceptable integrity rating, i.e., as re-specified, the extent to which the measure retains the originally stated intention of the measure.
- Percentage of test sites reporting an acceptable face validity rating, i.e., the extent to which the measure appears to capture the single aspect of care or healthcare quality as intended, and the measure as specified is able to differentiate quality performance across providers.

Step 11: Support the COR/GTL in the submission of testing results to NQF

If the measure(s) will be submitted to NQF for endorsement, the measure contractor assists the COR/GTL, as directed, in the completion of the measure submission form that includes results of measure testing. The information documented in the MIF should be used to complete the NQF submission. Measure contractors also provide additional information as needed, and are available to discuss testing results with NQF throughout the endorsement process.

13.5 Testing and Measure Evaluation Criteria

Step four of the measure testing procedure describes the analysis of testing results. They are used to demonstrate a measure's alignment with the measure evaluation criteria. Because testing is often an iterative process, both alpha and beta testing findings may provide information that address measure evaluation criteria.

- Alpha testing often supplies information that demonstrates the feasibility of the measure's implementation.
- The findings from one or more beta tests are often used to demonstrate scientific acceptability and usability, as well as augment previously obtained information on the importance and feasibility of the measure.

Application of the testing results to each of the four measurement areas (importance, scientific acceptability, usability, and feasibility) is discussed below.

Importance

Information from testing often provides additional empirical evidence to support prior judgments of a measure's importance generated earlier during the measure development process. In particular, beta testing results can be used to demonstrate that a measure assesses an area with substantial

opportunities for improvement. Testing can also help demonstrate that the measure addresses a high-impact or meaningful aspect of health care. Examples of empirical evidence for importance or improvement opportunities derived from testing data include:

- Quantifying the frequency or cost of measured events to demonstrate that rare or low-cost events are not being measured.
- Identification of substantial variation among comparison groups, or suboptimal performance for a large proportion of the groups.
- Demonstrating that methods for scoring and analysis of the measure allow for identification of statistically significant and practically/clinically meaningful differences in performance.
- Showing disparities in care related to race, ethnicity, gender, income, or other classifiers.
- Evidence that a measured structure is associated with consistent delivery of effective processes or access that lead to improved outcomes.

Reported data to support the importance of a measure may include:

- Descriptive statistics such as means, medians, standard deviations, confidence intervals for proportions, and percentiles to demonstrate the existence of gaps or disparities.
- Analyses to quantify the amount of variation due to comparison groups such as rural versus urban (e.g., r^2) or providers or hospitals (e.g., intra-class correlation).

Scientific Acceptability

With respect to CMS and NQF review for endorsement, scientific acceptability of a measure refers to the extent to which the measure produces reliable and valid results about the intended area of measurement. These qualities determine whether the measure can be used to draw reasonable conclusions about care in a given domain. Because many *measure scores* are comprised of patient-level *data elements* (e.g., blood pressure, lab values, medication, or surgical procedures) that are aggregated at the comparison group level (e.g., hospital, nursing home, or physician), evidence of *reliability* and *validity* is often needed for both the measure score and measure elements, and the measure developer should ensure both are addressed. Some examples of common measure testing and reporting errors are shown in Table 13-2.

Table 13-2 Examples of Common Errors in Measure Testing and Reporting

Common Error	Description
Reporting only descriptive statistics	Reporting the calculation of measure scores and the measure descriptive statistics, which demonstrate data are available and can be analyzed, but not reporting evidence of reliability or validity.
Inadequate testing of adapted measures	When adapting a measure, assuming that because a similar measure is already in use, it does not require testing to obtain empirical evidence of reliability and validity.

Common Error	Description
Inadequate scientific acceptability evidence for measure elements	Assuming that a measure's elements (e.g., diagnosis codes, EHR fields) that are in common use do not require testing or other evidence for reliability and validity within the context of the new measure specifications (e.g., new population, new setting).
Inadequate evidence for exclusions	Failing to report analyses justifying exclusions or demonstrating reliability across different methods of data collection.

Since reliability and validity are not all-or-none properties, many issues may need to be addressed to supply adequate evidence of scientific acceptability. However, the complexity of different health care environments, data sources, and sampling constraints often preclude ideal testing condition. As such, judgments about a measure's acceptability are often a matter of degree. Therefore, determination of adequate measure reliability and validity is always based upon the review of the testing data by qualified experts; it is assumed that a measure developer will contract or employ experienced methodologists, statisticians, and subject matter experts to select testing that is appropriate and feasible for the measure(s) under consideration, and ensure demonstration of measure reliability and validity.

While not replacing the expert judgment of the measure development team, the following subsections describe the general considerations for evaluating reliability and validity of both a measure score and its component elements.

Reliability

Reliability testing demonstrates that measure results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period.

Measure data elements versus measure score

Because many measures are comprised of multiple data elements, reliability testing ideally applies to both the data elements comprising the measure and the computed measure score. However, for measures that rely on many data elements, testing of the individual data elements is sometimes only conducted for critical elements that contribute most to the computed measure score, rather than all of the data elements. Similarly, commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission) are also occasionally excluded from reliability testing, although some mistakes can happen there, too. Prior evidence of validity of the data elements can also be used in place of reliability testing of the measure's elements since the two concepts are both mathematically and conceptually related such that sufficient reliability is required to have sufficient validity. That is why established sufficiency of validity also establishes reliability.

This flexibility in the reliability testing of data elements contrasts with assessment of the measure score. The measure score under development should always be assessed for reliability using data derived from testing.

Types of reliability

Depending upon the complexity of the measure specifications, one or more types of reliability may need to be assessed. Several general classes of reliability testing are shown below:

- Inter-rater (inter-abstractor) reliability—Assesses the extent to which ratings from two or more observers are congruent with each other when rating the same information (often using the same methods or instruments). It is often employed to assess reliability of data elements used in exclusion specifications, as well as the calculation of measure scores when review or abstraction is required by the measure. The extent of inter-rater/abstractor reliability can be quantitatively summarized, and concordance rates and Cohen's Kappa with confidence intervals are acceptable statistics to describe inter-rater/abstractor reliability. More recent analytic approaches are also available that involve calculation of intraclass correlations for ratings on a scale, where variation between raters is quantified for raters randomly selected to rate each occurrence.
- Form equivalence reliability (sometimes called parallel-forms reliability)—Assesses the extent to which multiple formats or versions of a test yield the same results. It is often used when testing comparability of results across more than one method of data collection or across automated data extraction from different data sources. It may be quantified using a coefficient of equivalence, where a correlation between the forms is calculated. As part of the analysis, reasons for discrepancies between methods (i.e., mode effects) should also be investigated and documented (e.g., when the results from a telephone survey are different from the results when the same survey is mailed).
- Temporal reliability (sometimes called test-retest reliability)—Assesses the extent to which a measurement instrument elicits the same response from the same respondent across two measurement time periods. The coefficient of stability may be used to quantify the association for the two measurement occasions. It is generally used when assessing information that is not expected to change over a short or medium interval of time. It is not appropriate for re-measurement of disease symptoms or intermediate outcomes that are expected to follow a given trajectory for improvement or deterioration. When used, a rationale for expecting stability (rather than change) over the time period should also be given.
- Internal consistency reliability—Testing of a multiple item test or survey assesses the extent that the items designed to measure a given construct are inter-correlated. [1] It is often used when developing multiple survey items that assess a single construct. Other internal consistency analysis approaches may involve the use of exploratory or confirmatory factor analysis.
- Other approaches to reliability—Across each type of reliability estimation described above, the shared objective is to ensure replication of measurements or decisions. In terms of comparisons

[1] Cronbach's alpha has been used to evaluate internal consistency reliability for several decades. Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*.16:297–334.

of groups, reliability can be extended to assess stability of the relative positions of different groups or the determination of significant differences between groups. These types of assessments address the proportion of variation in the measure attributable to the group (i.e., true differences or "signal") relative to the variation in the measure due to other factors (including chance variation or "noise"). Measures with a relatively high proportion of signal variance are considered reliable because of their power for discriminating among providers and the repeatability of group-level differences across samples. Provided that the number of observations within groups is sufficiently large, these questions can be partially addressed using methods such as analysis of variance (ANOVA), calculation of intraclass correlation coefficients, estimation of variance components within a hierarchical mixed (random-effects) model, or bootstrapping simulations. Changes in group ranking across multiple measurements may also add to an understanding of the stability of group-level measurement.

Validity

In measure development, the term validity has a particular application known as test validity. Test validity refers to the degree to which evidence, clinical judgment, and theory support the interpretations of a measure score. Stated more simply, test validity indicates the ability of a measure to record or quantify what it purports to measure; it represents the intersection of intent (i.e., what we are trying to assess) and process (i.e., how we actually assess it).

Types of validity

Validity testing of a measure score can be assessed in many different ways. While some view all types of validity as a special case of construct validity, researchers commonly reference the following types of validity separately: construct validity, discriminant validity, predictive validity, convergent validity, criterion validity, and face validity.[2]

- Construct validity—which refers to the extent to which the measure actually quantifies what the theory says it should. Construct validity evidence often involves empirical and theoretical support for the interpretation of the construct. Evidence may include statistical analyses such as confirmatory factor analysis of measure elements to ensure they cohere and represent a single construct.
- Discriminant validity/contrasted groups—which examines the degree to which the measure score diverges from other measures to which it theoretically should not be similar. It may also be demonstrated by assessing variation across multiple comparison groups (e.g., health care organizations, hospitals) to show that the measure can distinguish between disparate groups that it should theoretically be able to distinguish.
- Predictive validity—which refers to the ability of measure scores to predict scores on other related measures at some point in the future, particularly if these scores predict a subsequent

[2]Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

patient-level outcome of undisputed importance, such as death or permanent disability. Predictive validity also refers to scores on the same measure for other groups at the same point in time.

- Convergent validity—which refers to the degree to which multiple measures/indicators of a single underlying concept are interrelated. Examples include measurement of the correlations between a measure score and other indicators of processes related to the target outcome.
- Reference strategy/Criterion validity—which refers to verification of data elements against some reference criterion determined to be valid (the gold standard). Examples include verification of data elements obtained through automated search strategies of electronic health records compared against manual review of the same medical records.
- Face validity—which is the extent to which a measure appears to reflect that which it is supposed to measure "at face value." It is a subjective assessment by experts about whether the measure reflects what it is intended to assess. Face validity for a CMS quality measure may be adequate if accomplished through a systematic and transparent process, by a panel of identified experts, where formal rating of the validity is recorded and appropriately aggregated. The expert panel should explicitly address whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and poor quality. Because of the subjective nature of evaluating the face validity of a measure, special care should be taken to standardize and document the process used. NQF has recommended that a formal consensus process be used for the review of face validity such as a modified Delphi approach where participants systematically rate their agreement, and formal aggregating and consensus failure processes are followed. [3] This type of formal process can also be used when addressing whether specifications of the measure are consistent with medical evidence.

Measure data elements versus measure score

Validity testing applies to individual data elements used in a measure, as well as the computed measure score. Similar to reliability testing, validity testing is ideally conducted for both the data elements and the computed measure score to demonstrate that the measure data elements are correct—and that the measure score correctly reflects the targeted aspect of care.

Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Some examples of analysis of the validity of a measure data element include:

- Administrative data—Claims data where codes that are used to represent the primary clinical data (e.g., ICD, CPT) can be compared to manual abstraction from a sample of medical charts.

[3] Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties. (2011): The National Quality Forum—Task Force on Measure Testing.

- Standardized patient assessment instrument—Standardized information (e.g., MDS, OASIS, registry data) that is not abstracted, coded, or transcribed can be compared with "expert" assessor evaluation (conducted at approximately the same time) for a sample of patients.
- EHR clinical record information—EHR information extracted using automated processes based upon measure technical specifications can be compared to manual abstraction of the entire EHR (not just the fields specified by the measure). For measures that rely upon many data elements, testing may not necessarily be conducted for every single data element. Rather, testing may involve only critical data elements that contribute most to the computed measure score.

Prior evidence of reliability and validity for measure elements

When prior evidence of reliability or validity of the data elements comprising the measure exists, it can sometimes be used in place of testing of the measure's data elements. In contrast to a measure's data elements, while prior evidence can augment findings for a calculated measure score under development, it should always be assessed for reliability and validity using data derived from the beta test.

Prior evidence can include published or unpublished testing. NQF[4] provides the following guidance:

- Validity—Prior evidence of validity of data elements can be used for evidence provided the prior evidence uses the same data elements and data type as the measure under development, and it was obtained using a representative sample of sufficient size. Data elements that represent an existing standardized scale are also often excluded when a judgment is made that the validity of the scale has already been confirmed.
- Reliability—Separate reliability testing of the data elements is not required if validity testing is conducted on the data elements. If validity testing was not conducted, prior evidence of reliability of data elements can be used for evidence provided the prior evidence used the same data elements and same data type, and it was obtained using a representative sample of sufficient size.

Testing of exclusions/exceptions

When exclusions/exceptions are used in a measure's specifications, review of them based upon testing data is recommended. Review should include at a minimum:

- Evidence of sufficient frequency of occurrence of the exclusion/exception.
- Evidence that measure results are distorted without the exclusion/exception. For example, evidence that exclusions distort a measure may include variability of exclusions across comparison groups and sensitivity analyses of the measure score with and without the exclusion.

[4]Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties. (2011): The National Quality Forum—Task Force on Measure Testing. APPENDIX A, tables A-2, A-4.

- Evidence that measure elements (e.g., codes) used to identify exclusion/exception are valid.

When patient preference or other individual clinical judgment based upon unique patient conditions is allowed as an exception category, it requires additional review. In addition to analyses to demonstrate the strong impact of the exception on the measure, careful consideration should be given to whether patient preference can be influenced by provider interventions or whether it represents a clinical exception to eligibility. These measures should always be reported both with and without the exception, and the proportion of exceptions should be included for any group-level tabulations.

Risk adjustment and stratification

Beta testing should be used to evaluate an evidence-based risk adjustment strategy when the measure under development is an outcome measure. Risk adjustment is not needed when the measure being developed is a process measure.

Empirical evidence for the adequacy of risk adjustment or rationale that risk adjustment is not necessary to ensure fair comparisons must be provided.

Information should include analytic methods used, evidence of meaningful differences; if stratification is used, the stratification results should be included. Refer to the Risk Adjustment section for more information.

Usability

Formal usability testing is often not required, and a review of measure characteristics (e.g., descriptive statistics, dispersion of comparison groups) may be conducted by the TEP to determine usability of the measure for performance improvement and decision making. When more formal testing is required by CMS to assess the understandability and decision-making utility of the measure with respect to intended audiences (e.g., consumers, purchasers, providers, and policy makers), a variety of methods are available. These include:

- Focus groups.
- Structured interviews.
- Surveys of potential users.
- Formal cognitive testing.

These different methods often focus upon the discriminatory ability of the measure, and the meaning of the score as applied to evaluation of comparison groups or decision making. For example, a survey of potential users may be used to rate the clinical meaningfulness of the performance differences detectable by the measure, or to assess the congruence of decisions based upon measure summary data from a sample.

Feasibility

Testing can be used to assess measure feasibility to determine the extent to which the required data are available, retrievable without undue burden, and the extent to which they can be implemented for performance measurement. Some feasibility information may be obtained when assessing the validity

of the measure score or measure elements (e.g., quantifying the frequency of absent diagnosis codes when a target condition is present). Other feasibility information can be obtained through the use of systematic surveys (e.g., survey of physician practices tasked with extracting the information). More in-depth information may be gathered by conducting focus groups comprised of professionals who may be responsible for a measure's implementation.

Feasibility assessments should address the following:

- The availability of data (e.g., evidence that required data, including any exclusion criteria, is routinely generated and used in care delivery).
- The extent of missing data, measure susceptibility to inaccuracies, and the ability to audit data to detect problems.
- An estimate of the costs or burden of data collection and analysis.
- Any barriers encountered in implementing performance measure specifications, data abstraction, measure calculation, or performance reporting.
- The ability to collect information without violation of patient confidentiality, including circumstances where measures based on patient surveys or the small number of patients may comprise confidentiality.
- The identification of unintended consequences.

13.6 Special Considerations

Testing measures specified for use in EHRs (eMeasures)

As EHR systems become more widely available, additional clinical information about health care may also become widely available for use. However, there are a multitude of EHR systems in use today (particularly in the ambulatory care setting), and this diversity must be managed when measure specifications are developed for use across EHR systems. To address this issue, CMS requires new measures (or measures being retooled or re-specified for EHRs) to be specified using the health quality measures format (HQMF), which is a standard for representing a health quality measure (or clinical quality measure) as an electronic document. In alignment with this format, measure developers are expected to author eMeasures in the Measure Authoring Tool and specify measures using the Quality Data Model (QDM), both developed by NQF (refer to the eMeasure Specifications section). These requirements promote measures that are reliable and valid when extracted across diverse EHR systems, provided they are certified EHRs. However, they also raise new considerations when testing measures that include EHR specification accuracy, EHR validity testing, measure score and element testing, testing of retooled measures, and feasibility testing. These are discussed below.

Testing EHR specification accuracy

The HQMF requires that eMeasures are specified using Extensible Markup Language (XML), which ensures that the measure is described in machine-readable form. To help ensure the accuracy of these specifications, measure developers are expected to validate the content of the XML. This is often achieved using the following two methods:

- Syntactic validation—This method of accuracy validation ensures that the XML content follows (i.e., conforms to) specific constraints required by the HL7 HQMF Draft Standard for Trial Use and the XML patterns based on the NQF Quality Data Model. These quality-checking processes are built into the NQF-developed Measure Authoring Tool (MAT) application (refer to the eMeasure Specifications section). Alternatively, other methods to confirm XML conformance can be used (i.e., HL7 ISO Schematron).
  - The HL7 ISO Schematron is a possible mechanism for validating XML that is written outside the MAT, however, it may not include all of the components that are now built into the MAT. Additional resources for information- including technical specifications- on ISO Schematron may be found at http://standards. iso. org/ittf/PubliclyAvailableStandards/index. html.
- Narrative validation—A fully constructed eMeasure is an XML document that can be viewed in a standard web browser, when associated with an eMeasure rendering style sheet supplied by CMS. Exports of an eMeasure from the NQF Measure Authoring Tool include the eMeasure xml file, the eMeasure style sheet, and a standalone spreadsheet that contains all value sets referenced by the eMeasure. When rendered in a web browser, the eMeasure is in a "human-readable" format which allows the measure author to assess the extent to which the machine generated criteria correctly reflects the original measure criteria under development.

EHR validity testing

Ideally, certified EHRs will use clinical information recorded in discrete computer-readable fields, which potentially reduces errors in measure elements arising from manual abstraction or coding errors. However, even under these circumstances, measures need to be evaluated during measure testing. For example, a measure may be impacted by the complexity of the specifications, such that it is susceptible to differences in use of the data fields by different users or incorporates elements that are frequently entered into the wrong EHR fields. A measure may also be impacted simply due to small errors in the measure specifications, such as the code lists or exclusion logic. Given the evolving potential for standardized data extraction of complex clinical information, different options should be considered during measure testing. These include:

- Comparison to abstracted records—eMeasure developers should consider comparing measure scores and measure elements across multiple data sources where EHR data practices may vary. This generally requires extracting data from certified EHRs, and comparing this information to manual review of the entire EHR record for the same patient sample. Sampling from diverse EHRs and comparison groups is critical to assessment of the measure's susceptibility to differences in data entry practices or EHR applications. Standard assessment of the agreement between the two methods using a reference strategy/criterion validity approach is often sufficient to demonstrate comparability.
- Comparison to simulated QDM data set—Because efforts are ongoing to ensure that the multitude of EHRs in use today are capable of exporting information suitable for a measure specified using HQMF and QDM standards, measure developers may have difficulty obtaining a sample that is adequately representative of the different practice patterns and certified EHR

systems that will be in use when implementing a measure. To address this issue, validity testing can be augmented using a simulated data set (i.e., a test bed) that reflects standards for EHRs, and includes sample patient data representing the elements used by the measure specifications. Provided the data set reflects likely patient scenarios and is constructed using QDM elements, the output can be used to evaluate the eMeasure logic and coding specifications. This approach is sometimes referred to as semantic validation, whereby the formal criteria in an eMeasure are compared to a manual computation of the measure from the same test database.

Measure score and element testing

Although the NQF assessment criteria for validity and reliability of eMeasures allow testing at the level of either the data elements or the performance measure score, in the near term it is unlikely to conduct validity tests using performance measure scores.[5] Therefore, the current accepted approaches to measure testing at the data element level are testing of the measure in a simulated environment and testing the output of the EHR extraction compared to visual inspection as described above.[6] For the *measure score*, these requirements specify examination of reliability and validity of the electronically-extracted eMeasure through a comparison to findings from manual review and abstraction of the entire EHR record. For *measure data elements*, demonstration of validity is considered adequate if either:

- Adequate agreement is observed between data elements electronically extracted and data elements manually abstracted from the entire EHR is found; or
- Complete agreement is observed between the known values from a simulated QDM-compliant data set and the elements obtained when the eMeasure specifications are applied to the data set. NQF guidance further clarifies that reliability testing of measure elements may be supplanted by evidence of measure element validity.

De novo or newly developed measures, in addition to demonstrating validity with either of the above, should be assessed to confirm that the appropriate vocabulary codes and taxonomies have been used as well as the QDM data types.[7]

Reliability and validity testing for measures modified for EHR use (retooled)

Measures originally specified using data sources other than EHR (i.e., chart abstraction or administrative claims data) can be modified or retooled for use with EHRs. However, even if these measures were previously approved by CMS and show adequate reliability and validity, the eMeasure should be assessed for the following:

---

[5]National Quality Forum. *NQF Draft Requirements for eMeasure Review and Testing.* November 2011. Available at: http://www. qualityforum. org/. . . /Draft_Requirements_for_eMeasure_Review_and_Testing. aspx. Accessed August 1, 2012.

[6]National Quality Forum. *Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties*. Task Force on Measure Testing. 2011:50.

[7]National Quality Forum. *NQF Draft Requirements for eMeasure Review and Testing.* November 2011. Available at: http://www. qualityforum. org/. . . /Draft_Requirements_for_eMeasure_Review_and_Testing. aspx. Accessed August 1, 2012.

- Similarity to the originally approved specifications.
- Appropriately used QDM data types which represent the original measure specifications.
- Appropriately chosen codes and taxonomies (if they exist).[8]

Consequently, at a minimum, a crosswalk of the eMeasure specifications should demonstrate that they represent the original measure. NQF indicates that this evidence warrants a "moderate" rating for the reliability and validity of a retooled measure. NQF guidance reserves a "high" rating for a measure when, in addition to reliability and validity testing described above, testing also demonstrates that the re-specified measure produces similar results (within tolerable error limits) when compared to the previously approved measure that uses data sources other than EHR. Ideally, this may be achieved by obtaining representative samples of patients across EHRs and providers that allow calculation of the measure directly from EHRs, as well as calculation upon the same patients using previously implemented methods (e.g., administrative claims or chart abstraction). This allows comparison across the data sources to determine if the EHR implementation produces similar (or superior) findings relative to data sources and specifications already in use. Statistics indicating agreement between the methods and data sources often provide a succinct summary of the similarity of the retooled eMeasure relative to prior implementation of the measure that doesn't use EHRs. Further investigation of patients where the data sources or methods do not match may also provide evidence for the adequacy of the retooled measure when a review of the patient record allows a definitive judgment regarding the appropriate disposition of the patient with respect to the measure.

Feasibility of eMeasures

Prior to drafting initial eMeasure specifications, the measure contractor should consider the data elements necessary for the proposed measure, and conduct preliminary feasibility assessments (alpha testing) to confirm availability of the information within a structured format within the electronic health record. This will better ensure that a developed measure passes feasibility assessments during beta (field) testing.

When developing the feasibility testing plan, careful consideration should be made when determining the threshold for feasibility. Feasibility should reflect variability in measure complexity and be considered along a spectrum, as all barriers to feasibility are not equal. Feasibility will require, at a minimum: (1) Determination of which measure-specified data elements are typically captured in the EHR, (2) whether the EHRs can reliably extract the specified data, (3) the ability of the EHR to capture this data through customary workflow (4) identification of any barriers to implementation related to technical constraints of EHRs, and finally (5) whether the data captured in the EHR is captured in a form that is semantically aligned with the expectations of the quality measure. Note that when replacing previous measures that rely upon manual chart review, the time and costs related to additional data entry by clinicians needs to be carefully considered. When recruiting sites and vendors to test feasibility of a developed measure, contractors should consult with their COR/GTL regarding the

[8] Ibid.

Paperwork Reduction Act of 1995, which requires Office of Management and Budget (OMB) approval before requesting most types of information from the public.

Adapted measures

When adapting a measure for use in a new domain (e.g., new setting or population), construct the measure testing to detect important changes in the functionality or properties of the measure. The following changes should be reviewed, as applicable:

- Changes in the relative frequency of critical conditions used in original measure specifications when applied to a new setting/population (e.g., dramatic increase in the occurrence of exclusionary conditions).
- Change in the importance of the original measure in a new setting (e.g., an original measure addressing a highly prevalent condition may not show the same prevalence in a new setting, or evidence that large disparities or suboptimal care found using the original measure may not exist in the new setting/population).
- Changes in the location of data or the likelihood that data are missing (e.g., an original measure that uses an administrative data source for medications in the criteria specification, when applied to Medicare patients in an inpatient setting, may need to be modified to use medical record abstraction because Medicare Part A claims do not contain medication information due to bundling).
- Changes in the frequency of codes observed in stratified groups when the measure is applied to a new setting or subpopulation.
- Changes requiring recalibration of any existing risk adjustment model, and/or changes that make the use of the previous risk adjustment model inappropriate in the new setting/population.

When these or other important changes are observed, the measure developer should refine the measure.

Composite measures

A composite measure is a combination of two or more individual measures into a single measure that results in a single score. The use of composites creates unique issues associated with measure testing. For NQF endorsement, testing the measure composite score must be augmented by testing the individual components of the composite. However, this does not apply to measure components previously endorsed by the National Quality Forum (NQF), or for components of a scale/instrument that cannot be used independently of the total scale. Below are recommendations for testing a composite measure in support of submission to CMS for approval and NQF for endorsement.

Component reliability and validity testing

Examination of reliability and validity is recommended for both the composite and the components of the composite. Composite component must individually meet demonstrate adequate reliability and validity, but the composite measure as a whole also must meet these criteria.

Component coherence

Testing is recommended to determine if components comprising a composite adequately capture the construct it purports to measure. This may include determination that components adequately cohere together in the calculation of a latent construct (e.g., using confirmatory factor analysis to assess the construct) and various other correlation analyses such as the assessment of internal consistency reliability.

Appropriateness of aggregation methods

Because the method selected for combining components may influence interpretation of a composite measure result, testing should include examination of the appropriateness of the methods used to combine the components into an aggregate composite score. For example, testing of process measures may include examination of the adequacy of all-or-none, any-or-none, or opportunity-scoring approaches used to score the composite with respect to other outcomes of interest. For a composite outcome that uses differential weighting of the components, supported for the weighting scheme may involve regression of the components upon a “gold standard” outcome, if available.

In general, when a linear combination is used to score a composite, measure testing should be conducted to demonstrate that components contribute to variation in the overall composite score, or testing should attempt to justify why a minimally contributing component is included in the composite. Care should always be taken to address situations where the majority of variability in the composite is caused by a subset of the components. This can counteract prior attempts to demonstrate face validity or construct validity, as the composite may primarily reflect a single component of the composite (e.g., group differences on an emergency room composite measure may be largely determined by emergency department wait times because variability for this component may be large relative to the variability of all remaining composite components).

The appropriateness of methods to address component missing data when creating the composite score should also be assessed, and this analysis of missing component scores should support the specifications for scoring and handling missing component scores.

Feasibility and usability of composite components

Measure testing may also demonstrate that the measure can be consistently implemented across organizations by quantifying comparable variation for individual components, and demonstrate that the measure can be decomposed into its components at the group/organization level to facilitate transparency and understanding by the intended measure audience.

14. National Quality Forum Endorsement[1]

14.1 Introduction

To the extent feasible, the Centers for Medicare & Medicaid Services (CMS) uses measures that have been endorsed by the National Quality Forum (NQF) in the CMS public reporting and value-based purchasing programs. This section of the Blueprint explains actions and responsibilities of measure contractor regarding the measure submission process to NQF and also measure contractor's role during the NQF measure endorsement process. As mentioned in previous sections, NQF endorses measures only if they pass four measure evaluation criteria—Importance to Measure and Report, Scientific Acceptability of Measure Properties, Usability, and Feasibility. NQF requires measure harmonization as part of their endorsement and endorsement maintenance processes, placing after initial review of the four measure evaluation criteria. Since harmonization should be considered from the very beginning of measure development, CMS contractors are expected to consider harmonization as one of the core measure evaluation criteria. (Please refer to the Harmonization section for further information).

This section describes measure contractor's role based on NQF's Consensus Development Process (CDP), version 1.9. However, measure contractors are responsible for reviewing the latest version of the NQF measure endorsement processes and resources found on the NQF Web site at the following address:
http://www.qualityforum.org/Measuring_Performance/Consensus_Development_Process.aspx.

Measure contractors may note that in its efforts to improve upon the current measure endorsement process, NQF has designed a new 2-stage Consensus Development Process (CDP). NQF is conducting a pilot project to assess the feasibility of this new CDP process. If the pilot project is successful, NQF anticipates that the first measure reviews using the new process will begin sometime in 2013. A brief overview of the proposed 2-stage endorsement process is provided at the end of this section.

14.2 Procedure

Measure submission

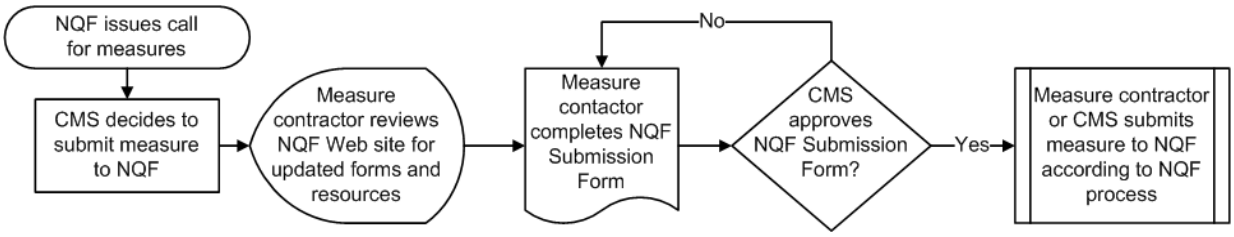
NQF's measure endorsement process is standardized in a regular cycle of topic-based measure evaluation. NQF endorses measures only as a part of a larger consensus development project to seek consensus standards (measures) for a given health care condition or practice. NQF follows a three-year schedule that outlines the review and endorsement of measures in 19 topic areas, such as cardiology, neurology, perinatal, and infectious disease. As the need arises, these topic areas may be revised to account for measures that may require a new or more appropriate topic area.

[1]The direction provided in this section is based on guidance from the National Quality Forum, and in some instances the verbiage remains unchanged to preserve the intent of the original documents.

Once a consensus development project is scheduled, at least two months before the start of that project, NQF issues a call for candidate standards (measures). During this time, NQF forms a Steering Committee for the project. A public call for nominations is posted on the NQF Web site.

The measure submission process is shown in Figure 14-1 and described in Steps 1 through 5 below.

Figure 14-1 Measure Submission



Step 1: NQF issues a call for measures

Before issuing a formal call for measures, NQF may post intent to call for measures on NQF's Web site. This an informal communication from NQF regarding the upcoming call for measures. In addition to the NQF Website, the information regarding intent to call for measures is distributed via email to NQF members and members of the general public who have signed up for public alerts. Developers who have measures they intend to submit for endorsement are asked to respond by email that includes:

- Title and description of each measure they intend to submit for evaluation.
- The name of the submitting organization and/or developer.
- The name, email address, and phone number of the contact person.

Approximately two weeks after the call for intent, NQF will issue a formal call for measures, requesting that all interested measure developers submit their relevant measures for consideration.

Measure contractors may note that at any point of time they may submit a notification to NQF that they have a measure that is ready, or almost ready, to submit. This "readiness to submit" notification increases NQF's awareness and knowledge of measures that are in the measure development pipeline. Measure contractors must obtain the approval of their Contracting Officer Representative/Government Task Leader (COR/GTL) before initiating the readiness to submit process. Completion of the "readiness to submit form" is not a formal submission of a CMS measure to NQF for consideration of endorsement. If a project is launched that accommodates the measure flagged for NQF, the submitter must still complete a full submission form during the call for measure for that particular project.

NQF also allows measure developers to submit measures online at any time, unrelated to a particular NQF project. Beginning the online submission form prior to an official call for measures allows the developers additional time to prepare and thoroughly review the submission form prior to submission. This may be useful when a large number of measures are being considered for submission by a single developer. Developers may also submit their measure at any time, however the measure will only be reviewed by NQF when there is a call or measures for that measure condition/domain.

Step 2: The COR/GTL decides to submit measure to NQF

The measure contractor should confirm the list of measures with the COR/GTL and begin preparing the measures for submission. The measure contractor and COR/GTL must inform the Measures Manager of upcoming measure submission. The measure contractor should review the NQF Web site (www.qualityforum.org) for updated forms and resources including directions on completing the online submission form. Every effort has been made to ensure that the MIF and Measure Justification form are aligned with the NQF Measure Submission Form to simplify the process of populating the data fields.

With the introduction of the EHR Incentive Program and Meaningful Use, there is a movement toward the development and/or retooling of measures specified for use in electronic health records (or eMeasures). The COR/GTL will provide guidance as to which eMeasures are candidates for NQF submission. These eMeasures, which are encoded in the Health Quality Measures Format (HQMF), are held to a different standard with respect to NQF submission. Measure contractors are responsible for monitoring NQF's eMeasure requirement policies prior to submission. At this time, the inclusion of eMeasures for all newly submitted measures is optional.[2]

Step 3: The measure contractor completes the NQF Measure Submission Form

Measure contractors must submit their measures via an online Measure Submission Form. The online form is available on the NQF Web site and allows users to:

- Gain secure access to the submission form from any location with an Internet connection.
- Save a draft version of the form and return to complete it at their convenience.
- Print a hard copy of the submission form for reference.

A user guide to the NQF Measure Submission Form is also available at http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx

When initiating an online measure submission form, the contractor may contact NQF and request additional users to access the online form. This allows the contractor to assign sections for the form to appropriate staff and facilitate internal review. The COR/GTL may also be listed as a user to facilitate ongoing and final review of the form. Contractors should inform their COR/GTL of this option. However, users must coordinate the timing at which they save their respective edits or else their edits could get over-written.

The CMS Measure Information Form (MIF) and Measure Justification Form (refer to the Measure Development section) have been aligned with the NQF Measure Submission Form. Both forms were designed to guide the measure contractor throughout the measure development process to gather the information needed for a successful NQF submission and organize it to minimize rework. The MIF and Measure Justification are CMS forms and present measures in a standardized way. They also provide a crosswalk to the fields in the NQF Measure Submission Form to facilitate online information entry.

[2]National Quality Forum: *Health Information Technology Advisory Committee (HITAC) Meeting Summary* May 22, 2012.

For contractors developing eMeasures, the HQMF document exported from the NQF-developed Measure Authoring Tool (MAT)—which includes a header and a body—should be used in lieu of the MIF. eMeasures are expected to be authored in the MAT, which is required to properly construct a CMS quality measure and the preferred format for NQF for endorsement. (Refer to the eMeasure Specifications section for further details.)

The measure contractor is responsible for completing the NQF Measure Submission Form and ensuring that the information is sufficient to meet NQF's requirements. The Measure Submission Form is the developer's presentation of the measure to the Steering Committee and others to demonstrate that the measure meets the criteria for endorsement. A Measure Submission Form is required for each measure submitted for endorsement. Listed below are a few tips for successful submissions:

- Answer every part of the NQF Measure Submission Form clearly and concisely.
- Provide substantive answers.
- Ensure that the form is complete enough to be read and understood as a stand-alone document.
- Attachments, references, and URLs are considered only supplementary and should include specific page numbers, table numbers, specific links, etc.
- Submit attachments or URLs as needed for long lists of codes or other data elements used in the measure, details of a risk adjustment model, and the calculation algorithm.
- Provide any pilot test data available, even if it does not satisfy NQF's entire pilot testing requirements.
- Identify all possible endorsement roadblocks in advance and address them in the Measure Submission Form.
- Document the rationale for all decisions made in the specifications.
- Document the rationale for all the measure exclusions.
- Discuss any controversies about the science behind the measure and why the measure was built as it was.
- Double check the document to ensure that no questions are left unanswered (i.e., no fields should be left blank and all questions should have a response).
- eMeasures must conform to the HL7 Health Quality Measures Format, have data criteria that are correctly mapped to the NQF Quality Data Model, and are expected to be authored in the NQF-developed Measure Authoring Tool.

There may be certain types of measures (e.g., composite measures) that do not have an online submission form. In these cases, NQF provides the form and instructions during the call for measures.

Contractors are encouraged to contact the Measures Manager for content questions while completing the online submission form. If, at any time, technical questions arise about the online submission form, or the need for technical support arises, NQF's Helpdesk is available at the email address: web-help@qualityforum.org. Questions about the content of or information required by the online submission form should be directed to the NQF project director whose name and contact information

appear on the project's Information page on the NQF Web site. In addition, Measures Manager will provide advice and support to the measure contractor as requested by the COR/GTL. Contractors are expected to have work closely with the Measures Manager throughout development—and provided all measure development deliverables—to ensure that no duplication of measure development occurs and to identify potential harmonization opportunities prior to NQF submission.

Step 4: The COR/GTL approves the NQF Measure Submission Form

The measure contractor will refer the completed Measure Submission Form to the COR/GTL for approval before submitting it to NQF.

The measure contractor should be aware that the COR/GTL may seek additional reviews of the completed Measure Submission Form before approving it. These reviews may come from the Measures Manager and other experts within CMS. Therefore measure contractors should account for that review period in their submission timeline.

The measure contractor will then make any necessary changes to obtain the COR/GTL's approval.

Step 5: The measure contractor or the COR/GTL submits the measure to NQF according to NQF processes

Once the COR/GTL has approved the Measure Submission Form, the measure contractor or the COR/GTL submits it to NQF using the online process.

Measure review by NQF

NQF follows a standardized measure review process to consider granting endorsement to the measure. NQF's current endorsement process is described below in Steps 6 through 8.

Step 6: Initial review by NQF staff

After NQF receives the Measure Submission Form, in addition to checking for completeness of the document, the NQF staff also performs an initial review to ensure that the measures submitted:

- Are in the public domain.
- Have a signed measure steward agreement.
- Are intended for BOTH quality improvement and public reporting.
- Have been tested or will be tested within 24 months.

Once this is done, the measures are accepted and are considered ready for review by the project Steering Committee.

Step 7: Steering Committee and Technical Advisory Panel Review

After the measures are accepted, the measures are reviewed by a Steering Committee that has been selected from nominations submitted during a public call for nominations. This Steering Committee oversees the work of NQF consensus development projects, offers expert advice, ensures that input is

obtained from relevant stakeholders, and makes recommendations to the NQF membership about measures that are proposed for endorsement.

A technical advisory panel may also be used depending on the technical expertise of the Steering Committee. Panel members are experts in their field and provide guidance to the Steering Committee around specific technical issues related to some or all of the measures under review.

The Steering Committee evaluates each submitted measure using the NQF Measure Evaluation Criteria. The Steering Committee evaluates each measure against the Importance criterion first. Only if the measure meets the Importance criterion it is evaluated against the other four criteria. After review, the Steering Committee recommends that a measure either continues through the consensus development process toward possible full or time-limited endorsement by NQF, or be returned to the measure steward and/or developer for further development and/or refinement. The measure contractor should be available to provide an overview and respond to questions during the Steering Committee meeting by attending in person or by teleconference. The COR/GTL, and a member from the Technical Expert Panel involved in measure development, may also attend the meeting to answer any questions that the Committee members may have on the measures.

Measures that are recommended for endorsement by the Steering Committee are then posted on the NQF Web site for public and NQF member comment. After the comment period, NQF members vote on the measures, and approved measures are then sent to the Consensus Standards Approval Committee (CSAC), a standing committee of the NQF Board of Directors. The CSAC is the governing body that makes endorsement decisions regarding national voluntary consensus standards. It has the most direct responsibility for overseeing the implementation of NQF's consensus development process.

Step 8: NQF determines type of endorsement appropriate for the measure

The CSAC reviews the recommendations of the Steering Committee and the results of the NQF member voting period. As during the Steering Committee review, the measure contractor, and COR/GTL (and possibly a member from the Technical Expert Panel) involved in measure development should attend the CSAC meeting in person or by teleconference to answer any questions that the CSAC members may have about the measures.

After its review of the measure, the CSAC may recommend one of the following:

- Grant full endorsement—requiring no further documentation. Per NQF, full endorsement is given to measures meeting NQF standard endorsement requirements.
- Grant a time-limited endorsement—for measures meeting all NQF requirements except adequacy of testing. Per NQF, time-limited endorsement status is granted to measures meeting all NQF endorsement requirements except field testing. This status is usually only granted for one year until testing is completed. On a very limited basis, the CSAC may decide that a candidate measure should be endorsed for a limited time period. This allows the NQF Board of Directors to make an endorsement before field testing is completed, when a measure otherwise meets or exceeds all other evaluation criteria. Prior to granting time-limited

endorsement status to a measure, the CSAC reviews and approves the measure developer's plan and timeline for field testing and provision of results to the CSAC.

- Measure stewards and developers may not request that a performance measure be considered for time-limited endorsement. This designation (originating from the steering committee and approved or disapproved by the CSAC) is made in most cases because a measure is determined to be urgently needed by the health care industry. Time-limited endorsement is not granted if a measure is complex (i.e., composite or requires risk-adjustment).

- Defer endorsement—This designation would be pending information from the measure steward.
- Decline to endorse the measure altogether—Per NQF, measures are not endorsed when they do not meet NQF endorsement requirements.

All of the CSAC's approval decisions are forwarded to the NQF Board of Directors for ratification. After a consensus standard (measure) has been formally endorsed by NQF, any interested party may file an appeal of the endorsement decision with the NQF Board of Directors. CSAC will make a recommendation to the NQF Board of Directors regarding the appeal. The Board of Directors will take action on an appeal within seven calendar days of its consultation with the CSAC.

Measure contractor's role during NQF review

The measure contractor supports CMS during the NQF review of the measure

During its review, NQF may have questions about the measure as submitted. These questions may come from NQF staff during initial measure submission or from the project Steering Committee. Questions may also arise during the NQF public comment period. The measure contractor proposes answers to the questions, and the COR/GTL reviews the proposed answers. Once the COR/GTL has approved the proposed answer, it is submitted to NQF. This may require consultation with the technical expert panel used by the contractor to develop or reevaluate the measures. For measures specified for electronic health records, the measure contractor, with support from the HQMF eMeasure subject matter expert, will work with NQF during eMeasure review.

To facilitate answering any questions, the measure contractor is encouraged to be actively involved in the NQF process as the measures are being considered. A member of the measure contactor's team who is prepared to explain and defend the measure should attend the Steering Committee and CSAC meetings as the measure is being discussed. The measure contractor will thus understand NQF's approach to the project in general and the measure contractor's measure(s) in particular. This active involvement provides a better position for the contractor to answer NQF's questions. During the discussions, the measure contractor should be prepared to defend the importance of the clinical topic, the scientific basis for each measure, the construction of the measure, and measure testing results.

The measure contractor makes changes suggested by NQF, with the COR/GTL's approval

During the NQF review, NQF may suggest changes to the measure to make it more acceptable, to harmonize with other measures, or both. If this occurs, the measure contractor may then consult with the technical expert panel used to develop or reevaluate the measures. With the COR/GTL's approval, the measure contractor makes the changes and provides the revised measure to NQF.

The measure contractor meets with the contractor's COR/GTL and the Measures Manager to review the NQF endorsement results and identify lessons learned

After NQF has completed its consensus development process (CDP), the measure contractor meets with the contractor's COR/GTL and the Measures Manager to discuss the results of the CDP, discuss why NQF came to their decision, identify lessons learned about both the NQF process and the CMS Measures Management System processes and discuss potential next steps.

Time-limited endorsed measures

On a very limited basis, NQF may grant a measure time-limited endorsement. This type of endorsement—generally granted based on urgent needs within the healthcare industry, rather than at the request of developers—stipulates that a measure must meet or exceed all evaluation criteria with the exception of field testing. If granted time-limited endorsement, measure developers must complete field testing within 12 months. Please refer to the Measure Testing section for further details on beta testing methods. Once the COR/GTL has approved the beta testing report, the measure contractor submits the documentation to NQF for review. Contractors are expected to be available to NQF during the review process. The COR/GTL must approve any changes.

NQF will notify the COR/GTL of the submission deadline for testing data, who will then inform the measure contractor—the Measure Manager may also provide this notification. However, measure contractors are responsible for tracking these dates and checking their NQF dashboard, as these notices are also available there.

The NQF Board may grant full endorsement if adequate field testing results are provided to and approved by CSAC before the time-limited endorsement period expires. Then that measure will be reevaluated at the end of its third year of endorsement along with the measures that were granted full endorsement from the beginning. After that, the measure will undergo regular maintenance every three years.

Measures with time-limited endorsement will lose that status and be de-endorsed if the measure contractor fails to provide testing results or seek an extension before the time-limited endorsement period expires. They will also be de-endorsed if the testing results fail to meet the NQF's measure evaluation criteria.

The NQF Time-Limited Endorsement Policy can be found at http://www.qualityforum.org/Measuring_Performance/Consensus_Development_Process/CSAC_Decision.aspx).

Expedited NQF review

In order to meet emerging national needs, NQF may expedite its review of certain measures for the CDP. NQF will obtain Consensus Standards Approval Committee (CSAC) approval before starting expedited review. The entire endorsement process for expedited review may be completed within four to five months. The COR/GTL and measure contractor will need to respond to requests quickly to meet this schedule.

NQF requires all of the following criteria be true prior to consideration by the CSAC for an expedited review**:**

- The measures under consideration must have been sufficiently tested and/or in widespread use. eMeasures will require only semantic testing using EHR simulated data sets (i.e., test beds)
- The scope of the project/measure set is relatively narrow.
- There is a time-sensitive legislative or regulatory mandate for the measures.

14.3 NQF 2-Stage Consensus Development Process (CDP)

Currently NQF is running a pilot project to assess feasibility of its newly designed 2-Stage CDP. It is anticipated that the first measure reviews using the new process would begin sometime in 2013.

Stage 1

In this stage, the contractors submit measure concepts to NQF. Contractors may ask NQF staff to review forms for accuracy, to ensure completion 30 days prior to the submission deadline. The measure concepts are then evaluated by the appropriate Steering Committee against the Importance to Measure and Report criterion. Per the NQF, stage 1 review will also help identify issues of harmonization and competing measures. Thereafter, measures concepts go to the CSAC and the Board for review. Once NQF has approved that the measure concepts met the importance criterion, they may then move into the second stage of the CDP. After that approval measure developers would have up to 18 months to bring the Stage 1 approved measure concepts back with specifications and testing information. If they are not submitted within 18 months, the measure concepts will require Stage 1 review again.

Another important thing for measure developers to note is that under the new process, NQF plans for the Steering Committees to meet for measure concept review every six months. Therefore, instead of waiting for the current three-year cycle, developers will have the opportunity to submit their measures up to twice a year. These committees, commissioned by NQF in 19 clinical and crosscutting topic areas, plan to schedule meetings throughout the year to evaluate the measure concepts.

Stage 2

In this second stage, measure developers submit fully tested and specified measures for review to gain NQF endorsement. The Steering Committee then evaluates the measures against the remaining criteria for NQF endorsement. After that, the measures go on to the CSAC and Board for final approval.

14.4 Measure Maintenance

Once NQF has endorsed a measure, the measure contractor supports the ongoing maintenance of the measure (if that support is part of the contractor's scope of work). Volume 2 of this Blueprint describes the standardized measure maintenance processes and the processes involved to maintain NQF Endorsement of the measures.

15. Measure Rollout

15.1 Introduction

This section describes the major tasks involved in the initial implementation (or rollout) of a measure at the national level. The process of rolling out measures varies significantly from one measure set to another depending on a number of factors, which may include, but is not limited to:

- Scope of measure implementation.
- Health care provider being measured.
- Data type and collection processes.
- Ultimate use of the measure (e.g., quality improvement, public reporting, pay-for-reporting, or value-based purchasing).
- Program into which the measure is being added.

The scope of measure implementation could be one of the following:

- A measure or measure set is implemented in a new program.
- A measure or measure set is added to an existing program.

Though the details of implementing measures may vary, the rollout processes generally include certain steps as defined within this section. The intensity or amount of effort involved in each of these tasks may vary and be affected by the factors listed above. When the Centers for Medicare & Medicaid Services (CMS) decide to start data collection at a national level, the measure is considered rolled out. Measure contractors should note that if approved by CMS, a measure may be implemented prior to complete rollout. For example, a measure may be used for facility-level quality improvement.

The work conducted in this section, as with all the sections of the Blueprint, will comply with the requirements of the Data Quality Act (DQA) as well as with the Department of Health and Human Services, *Guidelines for Ensuring the Quality of Information Disseminated to the Public*. (http://www.aspe.hhs.gov/infoquality/Guidelines/CMS-9-20.shtml).

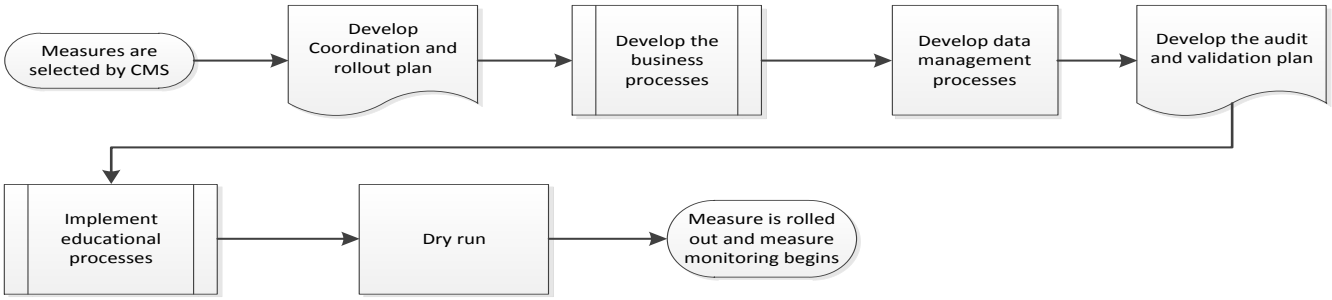
15.2 Deliverables

- Coordination and rollout plan
- Materials necessary to achieve approval from the Office of Management and Budget (OMB) for anything covered by the Paperwork Reduction Act (PRA)
- A data use agreement, if necessary
- Auditing and validation plan
- Appeal process for auditing and validation, if necessary
- Report on the results of any education processes that were conducted
- Report on the results of the dry run

15.3 Procedure

Figure 15-1 depicts the process of rolling out a measure approved by the Contracting Officer Representative/Government Task Leader (COR/GTL). A different set of tasks are involved for the ongoing use of the measure, which are provided in detail in the Measure Production and Monitoring section of Volume 2.

Figure 15-1 Overview of Measure Rollout Process



Once the COR/GTL has approved the use of a measure in a particular program, multiple tasks have to be carried out for rollout. These steps can be performed simultaneously whenever possible to achieve an efficient timeline. Review Figure 15-1 to identify which steps—detailed below—can be conducted simultaneously.

Step 1: Measures selected by CMS

Once the measures are developed, CMS selects measures for most of its programs by using the federal rulemaking process. This process is conducted in three phases:

1. Pre-rulemaking
2. Publishing the proposed rule for public comment
3. Issuing the final rule

Measure contractors will likely assist CMS with the rulemaking processes. For example, during the pre-rulemaking process contractors may need to provide information about their measures for presentation to the NQF-convened Measures Application Partnership. During the next phase, publishing the proposed rule for comment, the contractors will likely monitor the comments submitted on their measures and begin drafting responses for CMS. When the final rule is being issued, contractors may also be asked to provide additional documentation on their measures.

Step 2: Develop the coordination and rollout plan

The coordination and rollout plan includes the following key parts:

- Timeline for quality measure implementation
- Plan for stakeholder meetings and communication
- The anticipated business processes model
- The anticipated data management processes

- The audit and validation plan
- Plans for any necessary education processes

The coordination and rollout plan is referenced as the "implementation roadmap" in the Measure & Instrument Development and Support (MIDS) Umbrella Statement of Work (USOW).The anticipated business processes model and anticipated data management processes together represent the implementation algorithm reference in the MIDS USOW.

The COR/GTL will be responsible to inform and coordinate stakeholders during rollout—this may require substantial work. The measure contractor is responsible to coordinate and actively participate in stakeholder meetings, open door forums, or other means by which the public is informed of upcoming measure revisions. For this step, stakeholders may include, but are not limited to:

- State agencies
- Other CMS divisions
- Office of Information Services (OIS)
- Software vendors
- Providers and provider organizations

In addition to coordination with groups and individuals, the rollout plan may need to coordinate with other timelines. This may include the federal rulemaking process, the NQF measure review cycle, and other programs.

The intensity and types of communication will vary by program and measure. Some of the factors influencing the types of communication include the number of providers affected, the impact of the measures on the providers, the "newness" of the measure or program, and whether or not the rollout includes consumers. Some examples of communications strategies may include:

- Announcement to the Quality Improvement Organization (QIO) community by SDPS memo or to the ESRD network by email
- Presentations at conferences or scientific society meetings
- Publication of articles in peer-reviewed journals
- Publication in the Federal Register and the full rulemaking process
- National provider calls
- Press releases from CMS or CMS partners
- Notices in major newspapers across the country
- Town hall meetings with prominent CMS officials in various major cities
- Open door forums
- MedLearn articles
- Other processes as determined by the COR/GTL

The measure contractor must consider all of the variables listed above when developing the initial timeline for quality measure implementation. The timeline is then reviewed for approval by the COR/GTL.

The Paperwork Reduction Act (PRA) mandates that all federal government agencies must obtain approval from the Office of Management and Budget (OMB) before collection of information that will impose a burden on the general public. Measure contractors should be familiar with the PRA before implementing any process that involves the collection of new data. Contractors should consult with their COR/GTL regarding the PRA to confirm if OMB approval is required before requesting most types of information from the public. The full Act is available online at http://www.archives.gov/federal-register/laws/paperwork-reduction/. HHS also has an additional Web site with frequently asked questions and answers at http://www.hhs.gov/ocio/policy/collection/infocollectfaq.html.

Step 3: Implement the business processes

This step primarily applies to a new set of measures or a new use for an existing measure. The ultimate intended use of the measures will be a major factor in determining what occurs during the business modeling process. Examples of activities that can be conducted during this step include:

- Select measures for a new program—This may include a process for obtaining stakeholder input.[1]
- Develop the work processes and tools for data collection, rate calculation, and reporting.
- Develop the process for responding to questions about the measure.
- Identify which CMS divisions need to be involved to ensure that adequate resources are available when the measure is fully implemented.
- Determine any program rules, such as how eligibility for payment will be evaluated in a value-based purchasing or pay-for-reporting program.

Step 4: Implement the data management processes

The data management processes were created and tested during measure development. (Refer to the Measure Testing section). Like the measure, they must now be moved into a production environment.

The major tasks in this step include:

- Translate the algorithm into the production environment.
- Develop protocols and tools to receive data.
- Parallel processing of the data through the analysis program to ensure accuracy of the interpretation of the algorithm.
- Develop quality control processes.

Step 5: Develop the auditing and validation plan

The measure contractor will provide an audit and validation plan to the COR/GTL for approval before the measure is put into production.

[1]Refer to Section 3014 of the Accountable Care Act for ways in which the National Quality Forum-convened Measure Applications Partnership (MAP) provides multi-stakeholder input to HHS on the selection of performance measures for public reporting and payment reform programs. Also refer to the Paperwork Reduction Act regarding collection of new data.

There are several considerations when conducting audit and validation. The primary consideration is determining exactly what is being audited and/or validated, the measure or the data elements.

When auditing and validating data elements, the considerations include:

- Has the data been collected correctly? Were the algorithm and all auxiliary instructions followed? This is a particular concern for data that is abstracted from hard copy medical records where sampling methodologies and data hierarchies may be involved.
- Has the data been transmitted correctly? Are the standards for each data field maintained throughout the data transmission process? For example, abstraction instructions may require that dates be consistently expressed in mm/dd/yyyy format, but one or more mediating computer programs may employ yy/mm/dd formatting. If the calculation program relies on the first format, it may misread the second and adversely affect the provider's rate.
- Does the incoming data make sense? For example, a record might be suspect if it indicates a male receiving a hysterectomy or a female is diagnosed with prostate cancer.

When auditing and validating the measure itself, these considerations might include:

- If there are multiple databases used to calculate the rates, were they correctly linked?
- Was the sampling methodology correct?
- Were the data elements linked appropriately according to the measure specifications?
- Was the calculation algorithm programmed correctly?
- Do the measure results make sense? For example, rates greater than 100 percent may indicate an error in the calculation algorithm or in the calculation programming. Similarly, unexpectedly low rates may indicate a problem as well.

Step 6: Develop an appeals process

Before implementing a measure, CMS will determine if providers can appeal either the audit results or measure rates. The measure contractor may be required to assist in the development and design of these processes.

Step 7: Implement education processes

Providers will likely need to be educated on exactly what is being measured and how to interpret the results. For example, QIO networks may need to be informed on the measure and its meaning. For measures relying on abstracted data, abstractors must be trained to consistently identify qualifying cases and correct data. Methods for education include, but are not limited to:

- Conference calls and recordings of the calls
- Web-based presentations and recordings of the presentations
- Workshops at conferences or scientific society meetings
- Train-the-trainer events
- Other venues as determined by the COR/GTL

Step 8: Conduct the dry run

The dry run is the final stage of both measure testing and measure rollout. In the dry run, data is collected from all relevant providers across the country.

The purpose of the dry run is to finalize all methodologies related to case identification/selection, data collection (for measures using medical records data) and measurement calculation. It will verify that the measure design works as intended and begin to identify unintended consequences such as gaming or misrepresentation. The dry run also familiarizes relevant entities, such as CMS, QIOs, and the providers, with the reports. This provides the COR/GTL the opportunity to work with them to improve the usability of the reports before actual implementation and to identify and respond to questions and concerns. It also identifies any issues with the report production process so that the production processes can be improved to avoid problems when the measure is implemented.

Rates from a dry run are not publicly reported or used for payment or other reward systems, though the COR/GTL may decide to use them as the baseline measurement.

The dry run may not be a discrete step in the implementation of the measure. At the COR/GTL's direction, this step may be skipped—skipping this step means that the first round of data collection and results reporting may serve as the de facto dry run.

If problems arise during the dry run, those problems should be addressed and resolved before the measure is fully implemented.

Step 9: Submit reports

CMS may request reports summarizing the rollout processes. These may include reports:

- Describing the business processes.
- On the results of any education processes conducted.
- On results of the dry run, including but not limited to:
  - Analysis of the measure's success in meeting CMS' intentions for it.
  - Recommendations regarding:
    - Measure specifications.[2]
    - Business processes model.
    - Data management processes.
    - Audit and validation processes.
    - Educational processes for either data collectors or users of the measure results.

[2] A recommendation regarding changes to the measure specifications should clearly document the proposed changes and also address (at a minimum) whether the change is material or not, whether the change necessitates public comment or publication in the Federal Register, and whether the change affects other harmonized measures.

16. Glossary

1. **Access measure**—A measure that focuses on a patient or enrollee's attainment of timely and appropriate health care.

2. **Actionability** (Usability subcriterion)—Usefulness of the measure to multiple stakeholder groups in making decisions; i.e., the measure can be understood by:

   - Consumers to make health care decisions.
   - Health care organizations to improve quality.
   - Payers to adjust provider payments.
   - Individual providers to improve clinical decision-making.

   Also, the measure provides a distinctive or additive value to existing measures.

3. **Adaptability** (Usability subcriterion)—The extent to which the measure is adaptable to multiple populations or can be applied across various health care settings and includes information about specific situations under which the measure is applicable. Examples of adaptable measures include:

   - Smoking cessation counseling, regardless of reason for hospitalization.
   - Influenza immunization, regardless of setting.

4. **Adapted measures**—If an existing measure is changed to fit the current purpose or use, the measure is considered adapted. This may mean changes to the numerator or denominator, or changing a measure to meet the needs of a different care setting, data source, or population. Or, it may mean adding specifications to fit the current use.

5. **Adequacy of risk adjustment** (Scientific acceptability subcriterion)—Describes a risk adjustment model which reduces, removes, or clarifies the influences of confounding factors that differ among comparison groups.

6. **Adopted measures**—If a measure has the same numerator, denominator, data source, and care setting as its parent measure, and the only additional information that needs to be provided is particular to the measure's implementation use (such as data submission instructions), the measure is considered adopted.

7. **Alignment**—All aspects of a measure must be identical, and organizations using the measure must have agreements to continue maintaining the measure identically.

8. **Alpha testing** (also called formative testing)—

   - Size: Small Scale
   - Medical records: Less than 30 providers for measures that require data abstraction
   - Administrative data: Sample data set contains all of the data elements included in the measure data set

The purpose of alpha testing is to begin to assess the measure's feasibility, implementation barriers, and certain aspects of validity to determine whether the methods designed to collect the data or calculate the measure results could function as intended. Another purpose of alpha testing is to estimate the costs and burden of data collection and analysis.

9. **Attribution**—Assignment of the results of a measure to an individual, group, or organization responsible for the decisions, costs, and outcomes.[1]

10. **Audit**—A systematic inspection of records or accounts to verify their accuracy.

11. **Authoring tool**—an application that will enable measure developers to directly author measures using the Quality Data Set and automatically create a standards-compliant measure, thus avoiding the need for retooling. See Quality Data Set and Retooling.

12. **Beta testing** (also called field testing)—Size: Medium scale: Multi-site testing in a variety of appropriate settings with 50 to 100 health professionals, or 20 hospitals or large providers of different types, representing the full spectrum of the population being measured. The purpose of beta testing is to assess the measure's reliability to ensure that the specifications, data collection instructions, and computer programs are clear and concise. They should generate the same results regardless of when, where, or by whom data are collected and computer programs are run; continue to assess the measure's feasibility and implementation barriers and validity to determine if the refined methodologies designed to collect or extract the data and/or calculate the measure results function as intended; begin to identify unintended consequences, such as gaming or intentionally misrepresenting information; analyze the exclusions to assess their appropriateness and necessity; provide baseline performance data, including variation among providers and testing sites and evidence of a gap, such that there is opportunity for improvement; and, for outcome measures, assess the adequacy of the risk adjustment process.

13. **Bootstrap analysis**—In risk adjustment models, bootstrapping generally refers to estimating properties of a model estimate or the stability of an estimate by sampling from an approximating distribution. This is often accomplished by constructing many resamples of equal size from the observed dataset (for example, the development sample), where the resamples are smaller than the observed dataset. This technique allows estimation of the sample distribution of a statistic. It can also be used to construct hypothesis tests. In the case of a regression or logistic regression risk adjustment model, it can be used to provide additional guidance regarding the inclusion of risk factors in the model.

14. **Burden of data collection** (Feasibility subcriterion)—The extent to which a measure can be implemented without undue burden (financial or human) and in a manner that allows for auditing or verification of results. For example:

    - Data sources are accessible for the numerator, denominator, and exclusions;
    - Data sources contain the specified data;
    - Currently available data collection tools can be easily adapted;
    - Data can be collected from existing electronic data;

- Data can be generated during routine care delivery;
- Data are auditable.

15. **Business case**—A business case for a health care improvement intervention exists if the entity that invests in the intervention realizes a financial return on its investment in a reasonable time frame, using a reasonable rate of discounting. This may be realized as "bankable dollars" (profit), a reduction in losses for a given program or population, or avoided costs. In addition, a business case may exist if the investing entity believes that a positive indirect effect on organizational function and sustainability will accrue within a reasonable time frame.[2] The business case for a process measure relies on the financial return on the investment necessary to implement the intervention advocated by the measure. The business case for other types of measures relies on the financial return resulting from improving the quality of care indicated by the measure.

16. **Calculation algorithm**—An ordered sequence of data element retrieval and aggregation through which numerator and denominator events or continuous variable values are identified by a measure. Also referred to as the performance calculation.

17. **Cognitive testing**—Assessments of the cognitive capabilities of humans which pertain to the mental processes of perception, memory, judgment, and reasoning, as contrasted with emotional and volitional processes.

18. **Collection**—The highest possible level of the measure hierarchy. A collection may contain one or more sets, subsets, composites, and/or individual measures.

19. **Commercial interest**—A "commercial interest" as defined here, consists of any proprietary entity producing health care goods or services, with the exemption of non-profit or government organizations and non-healthcare-related companies.

20. **Comparable data**—The accuracy, reproducibility, risk-adjustability, and validity of the measure should not be affected if different systems use different sources for measures.[3] Data applied to a specific measure must be collected using similar methods and with a common definition throughout the population of interest.[4]

21. **Composite measure**—A combination of two or more individual measures in a single measure that results in a single score.[5]

22. **Concatenation**—A series of related events; to connect or link in a series or chain.[6]

23. **Concordance rate**—The proportion of a random sample of pairs that are concordant for a trait of interest.

24. **Conflict of interest**—Situation when an individual has an opportunity to affect measure contents that impact or serve an interest with which he/she has a relationship.

25. **Construct validity/discriminatory capability**—The extent to which performances measure demonstrates variation across multiple health care organizations; comparability assessment of a measure.

26. **Continuous Variable**—A measure score in which each individual value for the measure can fall anywhere along a continuous scale, and can be aggregated using a variety of methods such as the calculation of a mean or median (for example, mean number of minutes between presentation of chest pain to the time of administration of thrombolytics).

27. **Controllability** (Usability subcriterion)—The extent to which the measure is tied to health and medical care processes and outcomes that are under the control of the individual physician or other practitioner, provider organization, or health plan being measured.

28. **Convergent validity** (concurrent validity)—refers to the degree to which multiple indicators of a single underlying concept are correlated.

29. **Cost of care**—AQA defines cost of care as the total health care spending, including total resource use and unit price, by payer or consumer, for a health care service or group of health care services associated with a specified patient population, time period, and unit of clinical accountability.

30. **Cost/benefit** (Feasibility subcriterion)—The extent to which the benefit of measurement outweighs the financial and administrative burden of data collection, production of the measure, and implementation of the quality improvement interventions.

31. **Criteria**—Attributes or rules that serve as bases for evaluation, definition or classification of something; evaluation standards.

32. **c-statistic**—Used in the assessment of risk-adjusted models. The c-statistic is used in logistic regression with a dichotomous outcome (for example, alive/dead), and measures the area under a receiver operating characteristic (ROC) curve. The c-statistic indicates the ability of the model to discriminate between one event and the other. Random chance allows a model to discriminate randomly and c = 0.5. On the other hand, if the risk factors predict the outcome well, then discrimination goes up. The higher the c-statistic, the better the predictive power of the model.

33. **Data aggregation**—Refers to the combining of data from multiple sources for the purpose of generating performance information.

34. **Data element, critical**—Quality performance measures are based on many individual items of information. The data elements are often patient-level information on individual patients (for example, blood pressure, lab value, medication, surgical procedure, death). Testing at the data element level should include those elements that contribute most to the computed measure score, that is, account for identifying the greatest proportion of the target condition, event, or outcome being measured (numerator); the target population (denominator); population excluded (exclusions); and when applicable, risk factors with largest contribution to variability in outcome. Structural measures generally are based on organizational information rather than patient-level data.

35. **Data element, quality**—A quality data element is a single piece of information that is used in quality measures to describe part of the clinical care process, including both a clinical entity and its context of use (for example, diagnosis, active).

36. **Data flow attributes**—The fourth level of information in a Quality Data Model (QDM); Data flow attributes are descriptions of the authoritative source for the information that is required to represent any given quality data element. It includes the data source, recorder, setting and health record field.[7]

37. **Data sources**—The primary source document(s) used for data collection (for example, billing or administrative data, encounter form, enrollment forms, medical record).

38. **Denominator**—The lower part of a fraction used to calculate a rate, proportion, or ratio. The denominator is associated with a given patient population that may be counted as eligible to meet a measure's inclusion requirements.

39. **Denominator Exception**—Defined as allowable reason(s) for nonperformance of a quality measure for patients that meet the Denominator criteria and do not meet the Numerator criteria. Denominator Exceptions are the valid reasons for patients who are included in the denominator population, but for whom a process or outcome of care does not occur. These cases are removed from the denominator; however the number of patients with valid exceptions may still be reported. Exceptions allow for the exercise of clinical judgment. Allowable reasons fall into three general categories:

    - Medical reasons
    - Patients' reasons
    - System reasons

40. **De novo**—Literally meaning "from the beginning" or "anew," in the context of measures it refers to measures that have not been previously developed.[8]

41. **Denominator statement**—A statement that describes the population evaluated by the performance measure.

42. **Direct costs**—The dollar value of goods and services consumed as a result of illness and for which payment is made.[9]

43. **Discriminatory capability/construct validity**—The extent to which performances measure vary across multiple health care organizations; comparability assessment of a measure.

44. **Disparities in health care**—Health disparities are differences in health outcomes and their determinants between segments of the population, as defined by social, demographic, environmental and geographic attributes. [10]

45. **Dry run**—Full-scale measure testing involving all providers/practitioners representing the full spectrum of the population being measured. The purpose is to finalize all methodologies related to case identification/selection, data collection, and measurement calculation; and to quantify unintended consequences. The individual measure results are reported solely to verify

that the measure design works as intended. These results are not released to the public, nor are they the basis for any reward or sanction system. The results will be used to ensure that there are no unforeseen problems when the measure is implemented in the real world. In addition, the purpose for the dry run can be to establish evidence for a valid association between the process or structure and outcome of care where no such evidence exists for a measure; or to analyze certain statistical aspects of the measures as relates to a measure set. These results can be used to support selection of measures for public reporting and payment incentive programs, or for the development of composite measures.

46. **Economic case**—Describes the financial benefits of implementing a quality intervention at the provider or patient level, and within other aspects of society. A "business case" describes the financial impact of the intervention only on the investing entity.[11] An economic case describes the financial impact of the intervention to anyone other than the investing entity.

47. **Efficiency measure**—A measure that evaluates the relationship between a specific product (output) of the health care system and resources (inputs) used to create the product. For example, a provider in the health care system would be efficient if it was able to maximize output for a given set of inputs or to minimize inputs used to produce a given output. The Agency for Healthcare Research and Quality has developed a typology or analytical framework of efficiency measures: Perspective, outputs, and inputs.[12]

    The Institute of Medicine defines efficiency as avoiding waste, including waste of equipment, supplies and energy. To measure or assess efficiency, and ultimately value, associated with the care over the course of an episode of illness, the National Quality Forum has developed a framework to guide future and ongoing efforts in measuring efficiency in health care. The following constructs are essential to adequately assess the overall efficiency of the health care delivery system.[13]

    - *Quality of care* is a measure of performance on the six Institute of Medicine (IOM) specified health care aims: safety, timeliness, efficiency, equity, and patient-centeredness.
    - *Cost of care* is a measure the total health care spending, which includes total resource use and unit price(s), by payer or consumer, for a health care service or group of health care services, associated with a specified patient population, time period, and unit(s) of clinical accountability.
    - *Efficiency of care* is a measure of cost of care associated with a specific level of care. "Efficiency of care" is a measure of the relationship of the cost of care associated with a specific level of performance measured with respect to the other five IOM aims of quality.
    - *Value of care* is a measure of a specified stakeholder's preference-weighted assessment of a particular combination of quality and cost of care performance. Examples of stakeholders include individual patients, consumer organizations, payers, providers, governments, or societies.

48. **Efficiency of care**—a measure of cost of care associated with a specified level of health outcomes. AQA defines efficiency as a measure of cost of care associated with a specified level of quality of care.

49. **EHR**—Electronic health record (EHR) is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports. The EHR has the ability to generate a complete record of a clinical patient encounter—as well as supporting other care-related activities directly or indirectly via interface—including evidence-based decision support, quality management, and outcomes reporting.

50. **Electronic specifications**—Refers to measure specifications derived from EHRs and contain four main components:

    - Measure Overview/Description—contains the measure title, description, number, measurement period, measure steward, and other relevant information to the measure.
    - Measure Logic—contains population criteria and measure logic for numerator, denominator and exclusion categories. The measure logic contains the algorithm used to calculate performance.
    - Measure Code Lists—contains all of the codes pertaining to the measure.
    - Quality Data Set (QDS) elements—lists and describes each Quality Data Set (QDS) data element with the measure. See Quality Data Model.

51. **eMeasure**—An eMeasure is a health quality measure encoded in a health quality measure format (HQMF). See HQMF

52. **Empirical evidence**—Data or information resulting from studies and analyses of the data elements and/or scores for a *measure as specified*, unpublished or published.

53. **Epidemiological relevance** (Importance subcriterion)—Health problem/condition addressed by the measure, is a leading cause of mortality and/or morbidity, or is associated with a high incidence or prevalence rate for the population targeted by the measure.

54. **Exceptions**—See Denominator Exceptions.

55. **Exclusions**—See Denominator Exclusion and Numerator Exclusion.

56. **Expert consensus**—A parent term identifying recommendations formulated by one of several formal consensus development methods such as consensus development conference, Delphi method, and nominal group technique.

57. **Face validity**—The extent to which an empirical measurement appears to reflect that which it is supposed to "at face value." It is a subjective assessment by experts of whether the measure reflects the quality of care (for example, whether the proportion of patients with BP < 140/90 is a marker of quality.)

58. **Feasibility** (one of five major measure evaluation criteria)—Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement.

59. **Financial relationships**—Those relationships in which benefits by receiving a salary, royalty, intellectual property rights, consulting fee, honoraria, ownership interest (for example, stocks, stock options or other ownership interest, excluding diversified mutual funds), or other financial benefit. Financial benefits are usually associated with roles such as employment, management position, independent contractor (including contracted research), consulting, speaking and teaching, membership on advisory committees or review panels, board membership, and other activities from which remuneration is received, or expected. A minimal dollar amount for relationships to be significant has not been set. Inherent in any amount is the incentive to maintain or increase the value of the relationship. "Relevant" financial relationships in any amount occurring within the past 12 months that create a conflict of interest should be disclosed.

60. **Financial relevance** (Importance subcriterion)—Health problem/condition addressed by a measure is associated with a high annual cost or potential future medical costs for the population targeted by the measure.

61. **Gaming** (Part of Feasibility criterion, potential for unintended consequences subcriterion)—Includes limiting access to certain populations, neglecting care, or overuse of medications or services in order to ensure that the measure results are favorable.

62. **Gray literature**—Can include any documentary materials issued by government, academia, business, and industry such as technical reports, working papers, and conference proceedings that are unpublished or indexed commercially. As an example, contributors to the New York Academy of Medicine Grey Literature Web site include Agency for Healthcare Research and Quality (AHRQ), National Quality Forum (NQF), Centers for Disease Control and Prevention (CDC), Department of Health and Human Services (HHS), The Joint Commission (TJC), National Academy of Sciences, RAND, and RTI International.

63. **Guidelines**—Clinical practice guidelines are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances.[11]

64. **Harmonization** (Feasibility subcriterion)—Standardization of specific aspects of similar measures that, when different, do not increase the value or scientific strength of the measure. Examples of these aspects include: age ranges, denominator exclusions, data elements (for example, use of the same threshold for blood pressure in the same population), and codes.

65. **HITECH**—Health Information Technology for Economic and Clinical Health Act; A provision within the American Recovery and Reinvestment Act (ARRA) which authorizes incentive payments through Medicare and Medicaid to hospitals and clinicians towards meaningful use of electronic health records (EHRs). See Meaningful Use.

66. **HITEP**—Health Information Technology Expert Panel; An AHRQ-funded panel convened by NQF.

67. **HIT**—Health Information Technology allows comprehensive management of medical information and its secure exchange between health care consumers and providers.

68. **HQMF**—A standards-based representation of quality measures. A quality measure expressed in HQMF format is also referred to as an "eMeasure".

69. **Importance** (one of five major measure evaluation criteria)—extent to which the specific measure focus is important to making significant gains in health care quality (safety, timeliness, effectiveness, efficiency, equity, patient-centeredness) and improving health outcomes for a specific high-impact aspect of health care where there is variation in or overall poor performance.

70. **Included populations**—Detailed information describing the population(s) that the indicator intends to measure. Details could include such information as specific age groups, diagnoses, procedures, ICD-9-CM diagnostic and procedure codes, CPT codes, revenue codes, enrollment periods, insurance and health plan groups, etc.

71. **Indirect Costs**—Costs not related to care received by the patient. This may include lost income of patient or caregiver, absenteeism cost to employers, and in some cases such as relating to substance abuse, the cost associated with crime.

72. **Initial Patient Population**—The eligible group of patients that the performance measure is designed to address; usually focused on a specific disease process (for example, coronary artery disease, asthma). Details often include information based upon specific age groups, diagnoses, diagnostic and procedure codes, and enrollment periods. For example, a patient aged 18 years and older with a diagnosis of CAD who has at least 2 visits during the measurement period may represent a measure's initial patient population. All patients counted (for example as Numerator, as Denominator), are drawn from the Initial Patient Population.

73. **Institute of Medicine**

- Care needs
  - End of life care—Care related to those not expected to survive more than six months.
  - Getting better—Care related to acute illness or injury.
  - Living with illness—Care related to chronic or recurrent illness.
  - Staying healthy—Care related to healthy populations or the general health needs of non-healthy populations (for example, health promotion, disease prevention, risk factor assessment, early detection by screening and treatment of pre-symptomatic disease).
- Domains/dimensions
  - Effective—Providing services based on scientific knowledge to all who could benefit and refraining from providing services to those not likely to benefit (avoiding under use and overuse, respectively).
  - Efficient—Avoiding waste, including waste of equipment, supplies, ideas, and energy.
  - Equitable—Providing care that does not vary in quality because of personal characteristics such as gender, ethnicity, geographic location, and socioeconomic status.
  - Patient centered—Providing care that is respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions

- Safe—Avoiding injuries to patients from the care that is intended to help.
- Timely—Reducing waits and sometimes harmful delays for both those who receive and those who give care.

74. **Intellectual interest**—Intellectual interests may be present when the individual is a principle researcher/investigator in a study that serves as the basis for one or more to the potential performance measure under consideration.

75. **Intermediate Outcome**—The American Medical Association Physician Consortium for Performance Improvement (AMA PCPI) defines intermediate measures as measures that aim to meet specific thresholds of clinical care that have been shown to affect the desired health outcome (positively or adversely). Examples of intermediate outcome measures include blood pressure maintained at 140/90 or less, glycemic control, appropriate cholesterol levels reached.[14]

76. **Internal consistency reliability testing**—A multiple item test or survey to assess the extent that the items designed to measure a given construct are inter-correlated. Pertains to survey type measures and also pertains to the data elements used in measures constructed from patient assessment instruments.

77. **Inter-rater (inter-abstractor) reliability testing**—Assesses extent to which observation from two or more human observers are congruent with each other.

78. **Kappa statistics**—An index which compares the agreement against that which might be expected by chance. Kappa can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement), 0 (no agreement above that expected by chance) to -1 (complete disagreement).

79. **Leverage point**—A key state of health, clinical process, or event that has demonstrable effect on the health outcome. "Three considerations arise when evaluating leverage points: (1) the area being measured is an important contributing factor to the clinical or contextual process for the goal, (2) the area is one in which measurement and reporting is likely to stimulate improvement (through either selection or change), and (3) the purpose is to be selective rather than comprehensive."[15]

80. **Material change**—A material change is one that changes the specifications of an endorsed measure to affect the original measure's concept or logic, the intended meaning of the measure, or the strength of the measure relative to the measure evaluation criteria.

81. **Meaningful use**—A provision within the American Recovery and Reinvestment Act which authorizes the Centers for Medicare & Medicaid Services (CMS) to provide a reimbursement incentive for physician and hospital providers who are successful in becoming "meaningful users" of an electronic health record (EHR). These incentive payments begin in 2011 and gradually phase down. Starting in 2015, providers are expected to have adopted and be actively using an EHR in compliance with the "meaningful use" definition or they will be subject to financial penalties under Medicare.

82. **Measure Authoring Tool (MAT)**—An NQF-developed, publicly available, web-based tool for measure developers to create eMeasures; it should also reduce the time required to create new quality measures, and to convert existing paper-based measures in to EHR-readable format.[16]

83. **Measure evaluation criteria**—The CMS measure evaluation criteria serve as the basis for the measure development and evaluation processes used throughout the Measures Management System (MMS). The measure evaluation criteria consist of a number of subcriteria grouped under the following four main criteria: Importance, scientific acceptability, feasibility, and usability.

84. **Measure impact** (Importance subcriterion)—The human, social, and financial benefits from adhering to the measure have been quantified.

85. **Measure maintenance or evaluation**—the periodic and consistent reviewing and updating of performance measures to ensure currency with science, continued reliability, validity, feasibility, importance, and usability.

86. **Measure population**—Continuous variable measures do not have a Denominator, but instead define a Measure Population. To be in the measure population, a patient is in the larger Initial Patient Population appropriate to the measure set and is not excluded from the individual measure. Proportion and Ratio measures do not have a Measure Population, but instead define a Denominator.

87. **Measure score**—The numeric result that is computed by applying the measure specifications and scoring algorithm. The computed measure score represents an aggregation of all the appropriate patient-level data (for example, proportion of patients who died, average lab value attained) for the entity being measured (for example, hospital, health plan, home health agency, clinician, etc.). The measure specifications designate the entity that is being measured and to whom the measure score applies.

88. **Measure testing**—Empirical analysis to demonstrate the reliability and validity of the *measure as specified* including analysis of issues that pose threats to the validity of conclusions about quality of care such as exclusions, risk adjustment/stratification for outcome and resource use measures, methods to identify differences in performance, and comparability of data sources/methods.

89. **Measure Under Consideration**—A status referring to those measures that have not been finalized in previous rules and regulations, and that CMS is considering for the calendar year 2012.

90. **Measure, EHR**—An EHR measure is a health care quality measure specified for use with electronic health records; it is composed of data elements from the quality data set including code lists and measure logic, and can be translated to computer-readable specifications.

91. **Measure, untested**—Measure without empirical evidence of both reliability and validity. Untested measures are only eligible for time-limited endorsement if the conditions for considering time-limited endorsement are met.

92. **Measure**—A mechanism to assign a quantity to an attribute by comparison to a criterion.[17] A measure is the lowest level of measure hierarchy and may belong to a composite, subset, set, and/or collection. A distinction between the terms "quality indicator" and "quality measure" can be found in the report, "Overview of Risk Adjustment and Outcome, Measures for Home Health Agency OBQI Reports: Highlights of Current Approaches."[18]

93. **Medical record** (data source)—Data obtained from the records or documentation maintained on a patient in any health care setting (for example, hospital, home care, long-term care, practitioner office). Includes automated and paper medical record systems.

94. **Minor change**—A minor change does not change the process of data collection, aggregation, or calculation, nor does it change the intended meaning of the measure or the strength of the measure in terms of the measure evaluation criteria.

95. **Misrepresentation** (Part of Feasibility criterion, potential for unintended consequences subcriterion)—Refers to submission of incorrect information for the measure whether intentional or unintentional.

96. **Monetize**—Refers to the application of dollar amount (actual charges, standard price) to a unit of resource use. Monetizing resource use is an attempt to weight counts or resource units appropriately. For example, a frequency count of outpatient visits would give an equal count of one to both an office visit with an evaluation and an office visit with a procedure. Monetizing this would give a larger value to the office visit with a procedure.

97. **Morbidity**—The rate of incidence of a disease. The relative incidence of a particular disease morbidity. A diseased state or symptom (for example, lumbar puncture, if improperly performed, may be followed by a significant morbidity—Journal of the American Medical Association).

98. **Mortality**—The number of deaths in a given time or place. The proportion of deaths to population. "Death rate", also called "mortality rate."

99. **Multiple Chronic Conditions (MCC)—**Patients having two or more concurrent chronic conditions that collectively have an adverse effect on health status, function, or quality of life and that require complex healthcare management, decision-making, or coordination.[19]

100. **NPP**—National Priorities Partnership; NPP is a multi-stakeholder group organization convened by the National Quality Forum that offers consultative support to the Department of Health And Human Services on setting national priorities and goals for the HHS National Quality Strategy. The six NPP recommended priorities include:[20]

- *Health and Well-being*—This priority focuses on fostering health and wellness as well as national, state, and local systems of care that are fully invested in preventing disease, injury,

and disability, and that are reliable, effective, and proactive in helping all people reduce the risk and burden of disease.

- *Prevention and Treatment of Cardiovascular Disease*—This priority focuses on promoting the most effective prevention, treatment, and intervention practices for leading causes of mortality, beginning with cardiovascular disease.
- *Person and Family-Centered Care*—This priority focuses on honoring each individual patient and family, offering voice, control, choice, skills in self-care, and total transparency, and should adapt readily to individual and family circumstances, as well as differing cultures, languages and social background.
- *Patient Safety*—This priority focuses on reducing the risks of injury from care, aiming for "zero" harm wherever and whenever possible.
- *Effective Communication and Care Coordination*—This priority focuses on guiding patients and families through their health care experience, while respecting patient choice, offering physical and psychological supports, and encouraging strong relationships among patients and the health care professionals accountable for their care.
- *Affordable Care*—This priority focuses on assuring all patients access to affordable care that is delivered in a culturally and linguistically appropriate manner.

101. **National Quality Strategy Priority(-ies) (NQSP)**—A law within Section 3011 of the Affordable Care Act that seeks to increase access to high-quality, affordable health care for all Americans. The law requires the Secretary of the Department of Health and Human Services (HHS) to establish a National Strategy for Quality Improvement in Health Care (the National Quality Strategy) that establishes three aims:

- *Better Care*—To improve the overall quality of care by making health care more patient-centric, reliable, accessible and safe.
- *Healthy People/Healthy Communities*—To improve the health of the U.S. population by supporting proven interventions so that behavioral, social, and environmental determinants of health are addressed, in addition to delivering higher-quality care.
- *Affordable Care*—To reduce the cost of quality health care for individuals, families, employers and government.

Six priorities were established to aid in advancing the three aims:

- *Safety*—This priority focuses on making care safer by reducing harm caused in the delivery of care.
- *Person and Family Centered Care*—This priority focuses on ensuring that each person and his or her family members are engaged as partners in a care plan.
- *Communication and care coordination*—This priority focuses on promoting effective communication and coordination of care.
- *Effective prevention and treatment of illnesses*—This priority focuses on promoting the most effective prevention and treatment practices for the leading causes of mortality, starting with cardiovascular disease.
- *Best practices for healthy living*—This priority focuses on promoting wide use of best practices to enable healthy living.

- *Affordable care*—This priority focuses on promoting more affordable quality care for individuals, families, employers, and governments by developing and spreading new health care delivery models.

102. **Numerator**—The upper portion of a fraction used to calculate a rate, proportion, or ratio. A clinical action to be counted as meeting a measure's requirements (i.e., patients who received the particular service or obtained a particular outcome that is being measured).

103. **Numerator Exclusion**—Those patients who are included in the Initial Patient Population, who do not meet the measure numerator criteria, but who do meet the specific numerator exclusionary criteria. Numerator Exclusions are not considered to be part of a given measure's numerator.

104. **Numerator statement**—A statement that describes the clinical action that satisfies the conditions of the performance measure.

105. **Opportunity for improvement** (Importance subcriterion)—

- The measure is clearly related to a significant leverage point (a key aspect of a process) for achieving the intended goal; there is wide variation in quality, or quality is consistently substandard; or the measure is not at a level where rates can no longer rise.
- Substantive difference exists between recommended practices and actual practices.
- Evidence exists that interventions have been developed that purport to minimize this gap.
- Wide variation in performance exists among providers, subpopulations, or geographic regions.

106. **Outcome measure**—A measure that assesses the results of health care that are experienced by patients: Patients' clinical events; patients' recovery and health status; patients' experiences in the health system; and efficiency/cost.

107. **Parallel-forms reliability testing**—Assesses extent to which multiple formats or versions of a test yield the same results.

108. **Paperwork Reduction Act (PRA)**—mandates that all federal government agencies must obtain approval from the Office of Management and Budget (OMB) before collection of information that will impose a burden on the general public. Measure contractors should be familiar with the PRA before implementing any process that involves the collection of new data.

109. **Patient experience measure**—A measure that focuses on a patient or enrollee's report concerning observations of and participation in health care.

110. **Performance time frame**—A designated time frame within which the action described in a performance measure should be completed. This time frame is generally included in the measure description and may or may not coincide with the measure's data reporting frequency requirement. This can also be referred to as the numerator time window.

111. **Pilot testing**—Measure testing (sometimes referred to as pilot testing), is divided into two main types:

- Alpha testing (also called formative testing).
- Beta testing (also called field testing).

112. **Policy relevance** (Importance subcriterion)—Health problem/condition addressed by the measure is currently a policy priority. The measure is related to a national goal or a vulnerable population. For example:

- The measure is included in legislative mandates.
- The measure is included in lists of goals developed by HHS/CMS/other national bodies (such as IOM aims or NQF priority areas).
- The measure addresses disparities in health care quality or access related to race, ethnicity, age, socioeconomic status, income, region, gender, primary language, disability, or other classifications.
- Potential for unintended consequences (Feasibility subcriterion)—The extent to which a measure is vulnerable to gaming or misrepresentation. For example:
- Gaming includes limiting access to certain populations, neglecting care, or overutilization of medications or services to ensure that the measure results are favorable.
- Misrepresentation refers to submission of incorrect information for the measure whether intentional or unintentional.

113. **Predictive validity**—Ability of measure scores to predict scores on some other related valid measure. Predictive validity refers to the degree to which the operationalization can predict (or correlate) with other measures of the same construct that are measured at some time in the future.

114. **Precision of specifications** (Scientific acceptability of the measure properties subcriterion)—The extent to which specification details and key terms are precisely delineated and defined for each of the measure components, using clear language so there is no room for interpretation.

115. **Process measure**—A measure that focuses on a process which leads to a certain outcome, meaning that a scientific basis exists for believing that the process, when executed well, will increase the probability of achieving a desired outcome.

116. **Proportion**—A score derived by dividing the number of cases that meet a criterion for quality (the numerator) by the number of eligible cases within a given time frame (the denominator) where the numerator cases are a subset of the denominator cases (for example, percentage of eligible women with a mammogram performed in the last year).

117. **Protects confidentiality ("Feasibility" sub criterion)**—Patient confidentiality can be easily protected. For example:

- Data sources do not include individuals other than the population of interest.
- Arrangements have been made to meet HIPAA requirements regarding aggregation of small populations.

118. **Proxy variable**—In risk adjustment models, a proxy variable may not be directly of interest, but it can be used to obtain a measurement of a variable of interest. The proxy variable must be

correlated with the inferred value, but the correlation does not need to be perfect (i.e., $r = 1.0$).

119. **Public domain**—The realm embracing property rights that belong to the community at large, are unprotected by copyright or patent, and are subject to appropriation by anyone.[21]

120. **Quality data elements**—The third level of information in a quality data model (QDM); Quality data elements are a combination of a standard element and a quality data type that is used in quality measures to describe part of the clinical care process.[22]

121. **Quality data model (QDM)**—Developed by the National Quality Forum and formerly referred to as quality data set or QDS Model, a model of information that describes the clinical concepts in a standardized format so individuals (i.e., providers, researchers, measure developers) monitoring clinical performance and outcomes can communicate necessary quality improvement information clearly and concisely. The QDM describes the data elements and their context in four levels of information: standard elements, quality data types, quality data elements, and data flow attributes.[23]

122. **Quality data set**—See Quality data model.

123. **Quality data type**—The second level of information in a quality data model (QDM); Information that can be applied to a standard element to indicate the circumstance, or context, in which the standard element is used in a quality measure.[24]

124. **Quality Measure Set**—A unique grouping of performance measures carefully selected to provide, when viewed together, a general picture of the care provided in a given domain (for example, cardiovascular care, pregnancy).

125. **Quality Measurement and Health Assessment Group (QMHAG)**—Conducts Measures Priorities Planning to establish the quality measurement agenda for the Medicare program for the next five years. The Group is responsible for managing a number of quality measurement activities such as measure development, maintenance, implementation, and public reporting, which cover a number of health care service delivery settings such as hospitals, outpatient facilities, physician offices, nursing homes, and end-stage renal disease (ESRD) facilities.

126. **Quality measures and quality indicators**—A standard for measuring the performance and improvement of population health or of health plans, providers of services, and other clinicians in the delivery of health care services.

127. **Quality of care**—AQA defines quality of care as a measure of performance on IOM's six aims for health care: safety, timeliness, effectiveness, efficiency, equity, and patient centeredness.

128. **r^2 statistic**—Is frequently used to assess the predictive power of specific types of risk-adjusted models. Values for r^2 describe how well the outcome can be predicted based on the values of the risk factors or predictors.

129. **Ratio**—A score that may have a value of zero or greater that is derived by dividing a count of one type of data by a count of another type of data (for example, the number of patients with central lines who develop infection divided by the number of central line days).

130. **Rationale**—A brief statement describing the evidence base and/or intent for the measure that serves to guide interpretation of results.

131. **Receiver-operating characteristic (ROC) curve**—The graph that provides the c-statistic value is the ROC curve. The ROC curve graphs the predictive accuracy of a logistic regression model.

132. **Reference strategy (also known as a gold standard)**—Comparison of test results to an agreed upon "correct" result.

133. **Reliability** (Scientific acceptability of measure properties subcriterion)—

- *Measure reliability*: The results of the measure are reproducible a high proportion of the time when assessed in the same population (for example, the measure has high inter-rater reliability, no calculation errors, etc.).
- *Data element reliability*: The extent to which data elements used in the measure are free of identifiable errors (for example, coding errors).

134. **Reliability testing**—Empirical analysis of the *measure as specified* that demonstrate repeatability and reproducibility of the data elements in the same population in the same time period and/or the precision of the computed measure scores. Reliability testing focuses on random error in measurement and generally involves testing the agreement between repeated measurements of data elements (often referred to as inter-rater or inter-observer, which also applies to abstractors and coders) or the amount of error associated with the computed measure scores (signal versus noise).

135. **Reliability threats**—Refers to some aspects of the measure specifications or the specific topic of measurement that can affect reliability. Ambiguous measure specifications can result in unreliable measures. Small case volume or sample size, or rare events can affect the precision (reliability) of the measure score.

136. **Resource use measures**—Refers to broadly applicable and comparable measures of health services counts (in terms of units or dollars) applied to a population or event (broadly defined to include diagnoses, procedures, or encounters). A resource use measure counts the frequency of defined health system resources; some may further apply a dollar amount (for example, allowable charges, paid amounts, or standardized prices) to each unit of resource use—that is, monetize the health service or resource use units.

137. **Resource unit**—Refers to the resources used to provide care to a patient or population. Resource units are generally identified through claims data and measured in terms of dollars, but can also include resource not captured on a claim, for example, nursing hours.

138. **Retooling**—Conversion of measures from paper-based format to an electronic (eMeasure) format.

139. **Risk adjustment**—a statistical process used to identify and adjust for differences in patient characteristics (or risk factors) before comparing outcomes of care. The purpose of risk adjustment is to facilitate a fairer and more accurate comparison of outcomes of care across health care organizations or providers.

140. **Risk factor**—A variable associated with an increased/decreased risk of the outcome being measured. In measure development, it is often a patient-level characteristic, but it may also be associated with other levels in a model such as a hospital or geographic setting. Risk factors are correlational and not necessarily causal.

141. **Sample**—A subset of a population usually chosen in such a way that it can be taken to represent the population with respect to some characteristic

142. **Sampling frames**—The list of all cases potentially eligible for inclusion in the denominator, from which a more highly specified selection of cases will be made.

143. **Scientific acceptability of the measure properties (one of four major measure evaluation criteria)**—Extent to which the measure, *as specified*, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

144. **Scoring**—

- *Categorical variable*—A categorical variable groups items into pre-defined discrete, non-continuous classes (male, female), (board certified, not board certified). Categories may reflect a natural order, in which case they are called ordinal (cancer stage: I, II, III, or IV), (hospitals rankings: good, better, best).
- *Continuous variable*—A measure score in which each individual value for the measure can fall anywhere along a continuous scale (for example, mean time to thrombolytics which aggregates the time in minutes from a case presenting with chest pain to the time of administration of thrombolytics).
- *Frequency distribution*—A display of cases divided into mutually exclusive and contiguous groups according to a quality-related criterion.
- *Non-weighted score/composite/scale*—A combination of the values of several items into a single summary value for each case.
- *Rate*—A score derived by dividing the number of cases that meet a criterion for quality (the numerator) by the number of eligible cases within a given time frame (the denominator) where the numerator cases are a subset of the denominator cases (for example, percentage of eligible women with a mammogram performed in the last year).
- *Ratio*—A score that may have a value of zero or greater that is derived by dividing a count of one type of data by a count of another type of data (for example, the number of patients with central lines who develop infection divided by the number of central line days).
- *Weighted score/composite/scale*—A combination of the values of several items into a single summary value for each case where each item is differentially weighted (i.e., multiplied by an item-specific constant).

145. **Semantic validation**—A method of testing the validity of an eMeasure whereby the formal criteria in an eMeasure are compared to a manual computation of the measure from the same test database.

146. **Sensitivity—**Refers to the proportion of actual positives that are correctly identified as such (for example, the percentage of people with diabetes who are correctly identified as having diabetes).

147. **Set**—The second level of measure hierarchy. A set may include one or more subsets, composites, and/or individual measures.

148. **Social Case**—The description of the benefit to the individual patient or to society of the improved health status, regardless of cost presented to support a quality improvement effort or measure.[25]

149. **Specifications**—Measure instructions that address: data elements, data sources, point of data collection, timing and frequency of data collection, and reporting, specific instruments to be used (if appropriate) and implementation strategies.

150. **Specificity**—Refers to the proportion of negatives that are correctly identified—for example, the percentage of healthy people who are correctly identified as not having the condition. Perfect specificity would mean that the measure recognizes all actual negatives—for example, all healthy people will be recognized as healthy.

151. **Stakeholders**—Any person, group, or organization with an interest in, or who may be affected by, the activities of another organization.

152. **Standard deviation**—A measure of variability that indicates the dispersion, spread, or variation in a distribution.

153. **Standard Element**—The first level of information in a quality data model (QDM); Standard element is a clinical concept defined by a list of standard codes (for example, “diagnosis of heart failure” or “medication”). Each standard element contains a standard category (for example, diagnosis), a code set (ICD-10), and a code list (also known as a value set) of one or more codes.[26]

154. **Standardized price**—A pre-established uniform price for a service, typically based on historical price, replacement cost, or an analysis of completion in the market; removes variation in resource costs due to differences in negotiated prices.

155. **Stratification**—Refers to the division of a population or resource services into distinct, independent strata, or groups of similar data, enabling analysis of the specific subgroups. This type of adjustment can be used to show where disparities exist or where there is a need to expose differences in results.

156. **Strength of evidence base** (Scientific acceptability of the measure properties subcriterion)—Clinical measures follow evidence-based guidelines from medical specialty associations and relevant professional societies. In addition, supporting evidence includes consistent results

from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes or perception of care for various types of measures. For example:

- *Structure measures* use evidence that an association exists between the specific structural characteristic being measured and the outcomes of, or satisfaction with, care.
- *Process measures* use evidence that the measured clinical or administrative process leads to improved health or cost/benefit.
- *Outcome measures* are based on evidence that the outcome being measured can be impacted by one or more clinical interventions.
- *Access measures* use evidence that an association exists between the access measure and the outcomes of, or satisfaction with, care.
- *Patient experience measures* use evidence that an association exists between the measure of a patient's health care experience and the values and preferences of individuals/the public.

157. **Structural measure**—A measure that focuses on a feature of a health care organization or clinician relevant to its capacity to provide health care

158. **Subcriterion**—A constituent part or aspect of the four measure evaluation criteria, which should be considered when rendering a judgment on the acceptability of the quality measure being evaluated. Because the subcriteria address different aspects of the criteria, they should be considered individually, and then taken as part of the whole when rendering a decision on the overall assessment of the measure.

159. **Subset**—The third level of measure hierarchy. A subset may include one or more composites, and/or individual measures.

160. **Substitute variable**—Is also known as a proxy variable. See Proxy variable.

161. **Syntactic validation**—A method of accuracy validation which ensures that the Extensive Markup Language (XML) content of an eMeasure follows specific constraints required by the HL7 HQMF World Wide Web Consortium Schema and the NQF QDE XML pattern library.

162. **Systematic reviews**—Method for analyzing the evidence that includes a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research and to collect and analyze data from studies that are included with the review.

163. **Target population**—The numerator (cases) and denominator (population sample meeting specified criteria) of the measure.

164. **Temporal**—Refers to the time frame and related measure logic specified in a measure; occurring over a sequence of time or within a particular time.

165. **Testing**—The purpose of measure testing is to reveal the measure's strengths and limitations so that the limitations may be addressed and the measure can be refined and strengthened relative to the following measure evaluation criteria.

166. **Test-retest reliability testing**—Assesses extent to which a survey or measurement instrument elicits the same response from the same respondent across short intervals of time.

167. **Time window**—A time frame used to determine cases for inclusion in the denominator, numerator, or exclusions. The time frame includes an index event and period of time.

168. **Topped off**—A measure has reached a level where rates can no longer improve and so there is no opportunity for improvement.

169. **Usability (one of four major measure evaluation criteria**)—Extent to which intended audiences (for example, consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

170. **Validation**—Process (testing) to determine if a measure has the property of validity. The term validation is often used in reference to the data elements and is another term for validity testing of data elements. Validation also is used in reference to statistical risk models where model performance metrics are compared between two different samples of data called the development and validation samples.

171. **Validity** (Scientific acceptability of measure properties subcriterion)—

- *Measure validity*: The measure accurately represents the concept being evaluated and achieves the purpose for which it is intended (to measure quality). For example, the measure:
  - Clearly identifies the concept being evaluated (face validity).
  - Includes all necessary data elements, codes, and tables to detect a positive occurrence when one exists (construct validity).
  - Includes all necessary data sources to detect a positive occurrence when one exists (construct validity).
- *Data element validity*: The extent to which the information represented by the data element or code used in the measure reflects the actual concept or event intended. For example:
  - A medication code is used as a proxy for a diagnosis code.
  - Data element response categories include all values necessary to provide an accurate response.

172. **Validity testing**—Empirical analysis of the *measure as specified* that demonstrates that data are correct and/or conclusions about quality of care based on the computed measure score are correct. Validity testing focuses on systematic errors and bias. It involves testing agreement between the data elements obtained when implementing the measure as specified and data from another source of known accuracy. Validity of computed measure scores involves testing hypotheses of relationships between the computed measure scores as specified and other known measures of quality or conceptually-related aspects of quality. A variety of approaches can provide some evidence for validity. The specific terms and definitions used for validity may vary by discipline, including face, content, construct, criterion, concurrent, predictive, convergent, or discriminant validity. Therefore, the proposed conceptual relationship and test

should be described. The hypotheses and statistical analyses often are based on various correlations between measures or differences between groups known to vary in quality.

173. **Validity threats**—In addition to unreliability, some aspects of measure specifications and data can affect the validity of conclusions about quality. Potential threats include patients excluded from measurement; differences in patient mix for outcome and resource use measures; measure scores generated with multiple data sources/methods; and systematic missing or "incorrect" data (unintentional or intentional).

174. **Value of care**—Ambulatory Care Quality Alliance (AQA) defines value of care as a specified stakeholder's preference-weighted assessment of a particular combination of quality and cost of care performance. Examples of stakeholders include individual patients, consumer organizations, payers, providers, governments, or societies.

175. **Vulnerable population**—Groups of persons who may be compromised in their ability to give informed consent, who are frequently subjected to coercion in their decision making, or whose range of options is severely limited, making them vulnerable to health care quality problems. Examples include the frail elderly, minority populations, uninsured, etc.

176. **HL7** (Health Level 7)—A standards-developing organization that provides framework and standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services.

---

[1]Krumholtz, H., et al. Standard for measures: efficiency in health care. *Journal of the American College of Cardiology*. 2008;52 (number 18): p. 1522.
[2]Leatherman, S., Berwick, D., et al. The business case for quality: case studies and an analysis. *Health Affairs.* 2003; 22, (number 2): p. 18.
[3]National Committee for Quality Assurance. *IOM's Envisioning the National Health Care Quality Report,* 2001. Appendix E.
[4]National Research Council. *IOM's Envisioning the National Health Care Quality Report,* 2001. Appendix E.
[5]National Quality Forum, *Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety: Composite Measures, Draft Report*. July 2009, page 2.
[6]The American Heritage® Dictionary of the English Language, Fourth Edition copyright ©2000 by Houghton Mifflin Company. Updated in 2009. Available at: http://www.macmillandictionary.com/dictionary/american/concatenation. Accessed July 30, 2012.
[7]Truman B., Smith CK, Roy K, et al. CDC Health Disparities and Inequalities Report – United States, 2011. *MMWR Supplement* Vol. 2012;60:2. Available at: http://www.cdc.gov/mmwr/pdf/other/su6001.pdf Accessed July 30, 2012.
[8]National Quality Forum. *Quality Data Model Description*. Available at: http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx. Accessed July 30, 2012.
[9]Merriam-Webster dictionary; Available at: http://www.merriam-webster.com/dictionary/de%20novo
[10]Truman B., Smith CK, Roy K, et al. CDC Health Disparities and Inequalities Report – United States, 2011. *MMWR Supplement* Vol. 2012;60:2. Available at: http://www.cdc.gov/mmwr/pdf/other/su6001.pdf Accessed July 30, 2012.
[11]Leatherman, et al. p. 19.
[12]Agency for Healthcare Research and Quality. *Health Care Efficiency Measures*. Available at: http://www.ahrq.gov/qual/efficiency/. Accessed March 30, 2011.
[13]National Quality Forum (NQF). *Measurement Framework: Evaluating Efficiency Across Patient-Focused Episodes of Care.* Washington, DC: NQF; 2009.
[14]American Medical Association Physician Consortium for Performance Improvement. *Measures Development, Methodology, and Oversight Advisory Committee: Recommendations to PCPI Work Groups on Outcomes Measures. 2011*. Available at: http://www.ama-assn.org/resources/doc/cqi/outcome-measures-framework.pdf Accessed June 13, 2011.
[15]McGlynn EA. Selecting Measures of Quality and System Performance. *Medical Care.* 2003;41(suppl) 39-47.
[16]Available at: http://www.qualityforum.org/MAT/. Accessed August 1, 2012
[17]National Quality Measures Clearinghouse Glossary. Available at: http://www.qualitymeasures.ahrq.gov/about/glossary.aspx. Accessed June 13, 2011.
[18]Overview of Risk Adjustment and Outcome, Measures for Home Health Agency OBQI Reports: Highlights of Current Approaches
[19]National Quality Forum. *Multiple Chronic Conditions Measurement Framework*. Washington, DC: 2012. Available at: http://www.qualityforum.org/Publications/2012/05/MCC_Measurement_Framework_Final_Report.aspx. Accessed August 1, 2012.
[20]National Priorities Partnership. *Input to the Secretary of Health and Human Services on Priorities for the 2011 National Quality Strategy.* October 14, 2010. Available at: http://www.nationalprioritiespartnership.org/. Accessed June 5, 2012
[21]Merriam-Webster's Online Dictionary, Public Domain. Available at: http://www.merriam-webster.com/dictionary/public+domain. Accessed June 5, 2012.

[22]National Quality Forum. *Quality Data Model Description*. Available at: http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx. Accessed August 1, 2012.
[23]Ibid.
[24]Ibid.
[25]Leatherman, et al. p. 19.
[26]National Quality Forum. *Quality Data Model Description*. Available at: http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx. Accessed August 1, 2012.