

A Blueprint for the CMS Measures Management System

Version 9
Volume 2 of 2

August 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.....	1-5
1.6	Role of the Measure Contractor	1-6
1.7	Role of the Measure Contractor’s COR/GTL.....	1-7
1.8	Role of the Measures Manager	1-8
2.	Measure Evaluation During Maintenance Phase	2-1
2.1	Introduction	2-1
2.2	Deliverables.....	2-2
2.3	Discussion of Measure Evaluation Criteria and Subcriteria.....	2-2
2.4	Applying the Measure Evaluation Criteria	2-7
2.5	Using the Measure Evaluation Report and Measure Evaluation Guidance	2-9
2.6	Procedure.....	2-9
2.8	Overview of Appendices	2-12
3.	Harmonization During Maintenance Phase.....	3-1
3.1	Introduction	3-1
3.2	Procedure.....	3-2
4.	Measure Production and Monitoring	4-1
4.1	Introduction	4-1
4.2	Deliverables.....	4-2
4.3	Procedure.....	4-3
5.	Measure Maintenance.....	5-1
5.1	Introduction	5-1

5.2 Possible Outcomes.....	5-2
5.3 Measures Owned by Others	5-2
5.4 Criteria for Selection, Retirement, Removal and Suspension	5-3
6. Measure Update	6-1
6.1 Introduction	6-1
6.2 Possible Outcomes.....	6-1
6.3 Deliverables.....	6-2
6.4 Procedure.....	6-2
7. Comprehensive Reevaluation.....	7-1
7.1 Introduction	7-1
7.2 Possible Outcomes.....	7-1
7.3 Deliverables.....	7-2
7.4 Procedure.....	7-2
8. Ad Hoc Review	8-1
8.1 Introduction	8-1
8.2 Possible Outcomes.....	8-2
8.3 Deliverables.....	8-2
8.4 Procedure.....	8-3
9. Information Gathering During Maintenance Phase	9-1
9.1 Introduction	9-1
9.2 Deliverables.....	9-2
9.3 Procedure.....	9-2
10. Technical Specifications During Maintenance Phase	10-1
10.1 Introduction	10-1
10.2 Deliverables.....	10-3
10.3 Procedure.....	10-3
10.4 Special Considerations	10-14

11. eMeasure Specifications During Maintenance.....	11-1
11.1 Introduction	11-1
11.2 Deliverables.....	11-3
11.3 Health Quality Measures Format.....	11-3
11.4 National Quality Forum Quality Data Model	11-4
11.5 Building Block Approach to eMeasures	11-4
11.6 Measure Authoring Tool	11-5
11.7 Procedure.....	11-5
11.8 eMeasure Future and Next Steps	11-35
12. Technical Expert Panels During Maintenance Phase	12-1
12.1 Introduction	12-1
12.2 Deliverables.....	12-2
12.3 Procedure.....	12-2
13. Public Comment During Maintenance Phase.....	13-1
13.1 Introduction	13-1
13.2 Deliverables.....	13-1
13.3 Procedure.....	13-1
14. Measure Testing During Maintenance Phase.....	14-1
14.1 Introduction	14-1
14.2 Deliverables.....	14-1
14.3 Extent of Measure Testing.....	14-1
14.4 Testing For a Material Change	14-2
14.5 Alpha and Beta Testing	14-2
14.6 Procedure.....	14-6
14.7 Testing and Measure Evaluation Criteria.....	14-11
14.8 Special Considerations	14-19
15. National Quality Forum Endorsement Maintenance	15-1
15.1 Introduction	15-1

15.2 Procedure.....	15-1
15.3 NQF 2-Stage Consensus Development Process (CDP)	15-7
15.4 Conclusion.....	15-8
16. Glossary.....	16-1

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 2, Measure Maintenance	Description of Change
Section 1: Introduction	Added/Changed: <ul style="list-style-type: none"> Streamlined the “Background” section. Additional information about the process of updating the Blueprint. Added “Significant Changes in Version 9” section.
Section 2: Measure Evaluation During Maintenance Phase	Changed/Updated: <ul style="list-style-type: none"> Harmonization included as 5th measure evaluation criteria Usability criteria was updated to Usability and Use Guidance and tools for evaluating measures for Usability and Use (criteria and sub-criteria) were updated to align with NQF updated criteria.
Section 3: Harmonization During Maintenance Phase	New section within the Blueprint, information previously contained in other sections but consolidated due to importance of the topic.
Section 4: Measure Production and Monitoring	Added/Changed: <ul style="list-style-type: none"> Name changed to Production and Monitoring to reflect both tasks described in the section. Monitor tasks are now associated with each phase of production. Diagram depicting the process is simpler. Clarification regarding how feedback and questions from stakeholders are incorporated into the monitoring process Monitoring for unintended consequences is explicitly mentioned. More details regarding how information from monitoring can be used by CMS, such as in the pre-rulemaking process.
Section 5: Measure Maintenance	Added/Changed: <ul style="list-style-type: none"> Definition of “Suspend” to align with current CMS use. Criteria for Retiring a Measure. This section now includes CMS Measure Selection, Retirement, Removal and Suspension Criteria and is updated to reflect current definitions.

Section in Version 9.0, Volume 2, Measure Maintenance	Description of Change
Section 6: Measure Update	<p>Changed:</p> <ul style="list-style-type: none"> • Definition of “Suspend” to align with current CMS use. • Criteria for Retiring a Measure. This section now includes CMS Measure Selection, Retirement, Removal and Suspension Criteria and is updated to reflect current definitions.
Section 7: Comprehensive Reevaluation	<p>Changed:</p> <ul style="list-style-type: none"> • Definition of “Suspend” to align with current CMS use. • Criteria for Retiring a Measure. This section now includes CMS Measure Selection, Retirement, Removal and Suspension Criteria and is updated to reflect current definitions • Figure: Determining the appropriate outcome for reevaluation is changed to include the option of suspending a measure. <p>Added:</p> <ul style="list-style-type: none"> • PRA guidance regarding the solicitation of public comments during comprehensive reevaluation.
Section 8: Ad Hoc Review	Reformatted section for consistency.
Section 9: Information Gathering During Maintenance Phase	<p>Added/Changed:</p> <ul style="list-style-type: none"> • Added further guidance on the process of Information Gathering. • Reorganized and collapsed a few of the steps into a single step.
Section 10: Technical Specifications During Maintenance Phase	<p>Added/Changed:</p> <ul style="list-style-type: none"> • Section name changed from “Measure Specifications during Maintenance Phase” to “Technical Specifications During Maintenance Phase” for consistency between volumes 1 and 2. • Aligned the section with definitions in the eMeasure Specifications section. • Reordered Exclusions/Exceptions discussion to promote clarity. • Added language regarding ambiguous language regarding time references in measures. • Additional details regarding the NQF Endorsement Maintenance process. • Updated NQF ICD-10- CM/PCS timeline related to measure submission as a result of CMS’ proposed delay in implementation. • FAQs relocated to body of the section.
Section 11: eMeasure Specifications During Maintenance Phase	<p>Added:</p> <ul style="list-style-type: none"> • Section now provides specific guidance to eMeasure contractors regarding measure specification during the maintenance process. • Description- both verbal and visual- of the icon located throughout

Section in Version 9.0, Volume 2, Measure Maintenance	Description of Change
	<p>the Blueprint denoting special considerations for eMeasures.</p> <ul style="list-style-type: none"> • Figure 11-2 depicting the process of eMeasure specifications maintenance and transmission format. • Descriptive (verbal and visual) of the technical specifications maintenance process and factors influencing the development. • List of deliverables due to the COR/GTL at the completion of eMeasures Specifications maintenance process. • Metadata table (Table 11-1) - 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 11-C with detailed information on performance calculation of eMeasures. • Details regarding the NQF Endorsement Maintenance process specific to eMeasures. <p>Updated</p> <ul style="list-style-type: none"> • 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 12: Technical Expert Panels During Maintenance Phase	<p>Added/Changed:</p> <ul style="list-style-type: none"> • 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 13: Public Comment During Maintenance Phase	<ul style="list-style-type: none"> • Revised the MMS posting process to reflect the changes outlined by CMS. • Included language regarding the Paperwork Reduction Act (PRA).
Section 14: Measure testing During Maintenance Phase	<p>Added/Changed:</p> <ul style="list-style-type: none"> • Measure testing plan as a deliverable. • Table describing key features of alpha and beta testing. • Minor updates to clarify the reliability and validity sections, including

Section in Version 9.0, Volume 2, Measure Maintenance	Description of Change
	<p>updated references.</p> <ul style="list-style-type: none"> • Minor updates throughout the section to clarify that measure maintenance is completed by measure contractors, not measure developers. • Updated language referencing differences in testing between development and measurement stages. • 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 of eMeasures.
<p>Section 15: National Quality Forum Endorsement Maintenance</p>	<p>Added:</p> <ul style="list-style-type: none"> • Further guidance on measure contractor’s responsibility to track maintenance review due dates. • Information about obtaining permission for multiple users to access the measures submission form (to facilitate review prior to submission). • More information about relationship between CMS maintenance and endorsement processes. • NQF 2-step process.
<p>Section 16: Glossary</p>	<p>Additional terms reflective of the changes throughout version 9, volume 2 of the Blueprint.</p>



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/ngs/ngs2012annlrpt.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, health 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.

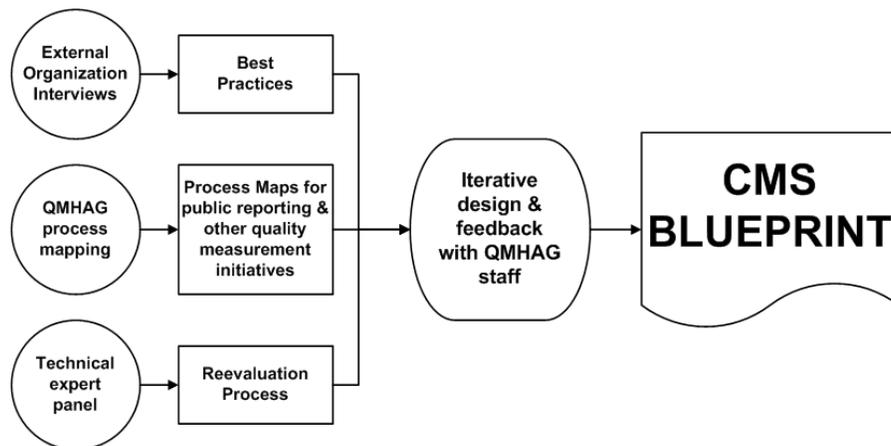
1.2 Development and Maintenance of the CMS Measures Management System Blueprint

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.

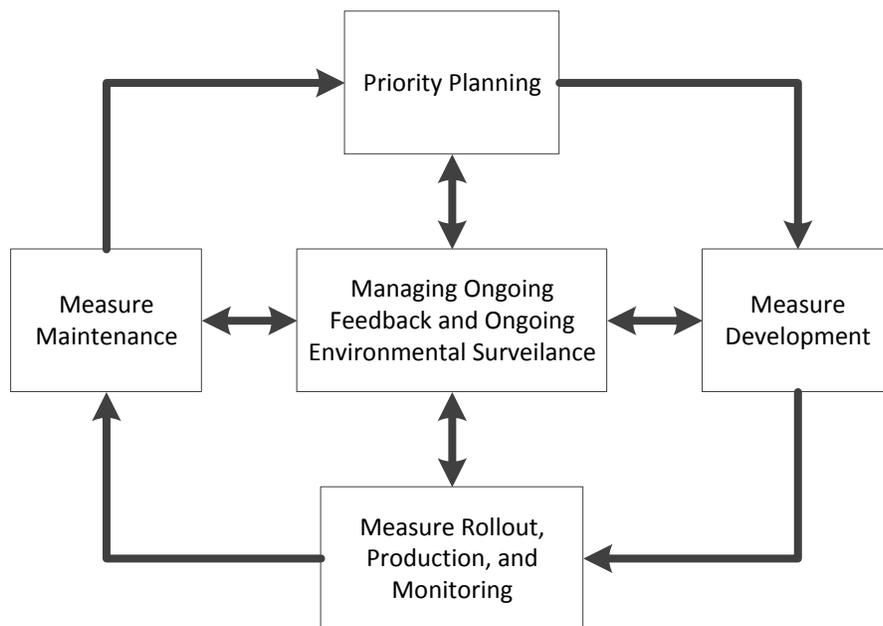
Two significant legislative Acts have recently affected quality measurement and led to Blueprint updates.

- ◆ 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 Evaluation During Maintenance Phase

2.1 Introduction

Similar to measure development, all of the hard work of the measure contractors, technical expert panel (TEP) members and stakeholders involved in measure maintenance is geared toward ensuring sound measures that can be used to drive health care quality improvement and inform consumer choice. In order to help the Centers for Medicare & Medicaid Services (CMS) ensure the continued soundness of its measures, the measure developer must provide strong evidence that the measure continues to add value to measurement programs and that it is constructed in a sound manner. This work also assists CMS to ensure that its measures maintain the endorsement of the National Quality Forum (NQF).

Each measure undergoes an annual update and a rigorous reevaluation every three years to assess its continued value and soundness based on the same set of standardized measure evaluation criteria used in measure development:

- ◆ Importance to measure and report
- ◆ Scientific acceptability of the measure properties
- ◆ Feasibility
- ◆ Usability and use
- ◆ 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 During Maintenance Phase section for further information).

In addition, to reevaluation based on the measure evaluation criteria, the contractor may be required to conduct an ad hoc review.

The measure contractor uses the measure evaluation process to update the measure justification and any changes to the technical specifications to demonstrate that:

- ◆ The aspects of care included in the specifications continued to be highly important to measure and report because the measurement results can supply meaningful information to consumers and health care providers that serves to drive significant improvements in health care quality and health outcomes where there is variation in performance or overall less-than-optimal performance.
- ◆ The data elements, codes, and parameters included in the specifications are the best ones to use to quantify the particular measure because they most accurately and clearly target the aspects of the measure 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 the best methodologies to use because they reflect a clear and accurate representation of the variation in the quality or efficiency of the care delivered or the variation in the health outcome of interest.

This section provides an overview of the Measure Evaluation Criteria and subcriteria. It also provides Measure Evaluation Guidance documents and a Measure Evaluation report template that can be used to document the measure's continued usefulness to CMS.

2.2 Deliverables

A separate measure evaluation report must be submitted to CMS for each measure at the following times:

- ◆ When recommending disposition of a measure after a comprehensive reevaluation.
- ◆ When recommending disposition of a measure after an ad hoc review.

2.3 Discussion of Measure Evaluation Criteria and Subcriteria

Importance to measure and report

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

High impact area

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 the opportunity for improvement. Impact should be systematically evaluated by updating the 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 case, economic case or social case. (Refer to Appendices in the Information Gathering During Maintenance Phase section 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. The topic area or measure focus should be evaluated relative to being 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, the step with the greatest effect on the desired outcome should be selected as the focus of measurement. For example, although assessment of 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 selection of a measure focus within a multi-step process that is consistent with the provider’s scope of practice.

Each of the three subcriteria in importance needs to be reevaluated during maintenance. The gap in care that was originally identified could be potentially eliminated after a measure has been in use for a number of years. As such, the measure may be no longer warranted for continued use and maintenance. In addition, the requirements for appropriate level of evidence to support a measure focus are influenced by the changing health care quality measurement environment. Even if no new studies have been published, the measure developer should reevaluate the evidence against the current criteria.

As mentioned above, the CMS quality measurement program is intended to drive health care quality improvement and inform consumer choice. When providers have done all they can to improve the quality of their care relative to a given measure, the measure is referred to as being “topped out” (i.e., achieving the highest rates that can reasonably be expected). A measure that has “topped out” can no longer be expected to aid in improving quality since the improvement has already been achieved to the extent the particular measure can drive it. Though there is no consensus on specific criteria for being

¹National 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 August 1, 2012.

“topped out,” CMS has proposed the following as guidelines for identifying measures that may have achieved their ultimate goal:

- ◆ “... high unvarying performance among [providers], as measures with very high performance among [providers] present little opportunity for improvement, and do not provide meaningful distinctions in performance for consumers.”²
- ◆ Measure performance rates showing that the 75th and 90th percentiles are statistically indistinguishable, showing the measure has topped out
- ◆ Measures exhibit a truncated coefficient of variation of less than (CV <) 0.10.³

For continued NQF endorsement, the importance criterion is considered “must pass” before the measure will be given any further consideration.

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 reliable and valid results about the quality of care when implemented. The subcriteria for both reliability and validity reflect this focus. Unlike new measure development, during measure maintenance the measure contractor has the advantage of experience with the measures to inform the assessment of this criterion.

Reliability

Evaluation of a measure’s reliability requires assessment of a measure’s specifications and empirical evidence of a 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. As with reliability, the validity assessment can be

²IPPS Final Rule published in the August 16, 2010 Federal Register.

³Proposed Rule for the Hospital Inpatient Value-Based Purchasing Program published in the Federal Register on January 7, 2011

aided by discussions found in the ongoing environmental scan conducted as a part of the measure monitoring process. Refer to the Measure Production and Monitoring section for details.

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, and 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 versus 75 percent) is clinically meaningful; or whether a statistically significant difference of \$25 in cost for an episode of care (e.g., \$5,000 versus \$5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

At reevaluation, the measure contractor has the advantage of seeing measure results from a much broader array of providers and patients than is usually seen during measure development. This enables measure contractors to demonstrate significance of differences between higher ranking and lower ranking providers, which is often missing during new measure development.

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.)

Reevaluation offers an opportunity to review the risk adjustment model to assess once again if it is the most appropriate model available.

Feasibility

This criterion evaluates the extent to which the required data are readily available, retrievable without undue burden, and implementable 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 electronic health records (EHRs) are more widely used, the measure contractors should be aware of the possibilities presented, and adapt their measures accordingly.

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. By definition, measures being reevaluated have an audit history that can be used to assess the measure's risks.

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. At reevaluation, outcome measures should be reviewed to determine whether they have been used to their maximum effect relative to such stimulation.

Harmonization

During maintenance phase, measures should be assessed for continued harmonization, keeping in mind that newer measures may have found better methodologies to achieve the same or similar quality improvement goals. 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, risk adjustment, exclusions, 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 During Maintenance Phase section for further details)

Measure contractors should be aware that if the measure is to be submitted to NQF for endorsement maintenance, and there are other measures essentially address the same target process, condition, event or outcome (numerator) and the same target population (denominator) the measures will be considered “competing measures.” Consult with the Measures Manager to review specifications to identify opportunities for further harmonization.

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.

If similar or related measures were identified the measure contractor should document why the measure being evaluated continues to have added value. 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.

2.4 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 endorsed by NQF, the measure must first pass all Importance subcriteria or else NQF will not evaluate it against the remaining criteria when NQF considers the measure for continued endorsement.
- ◆ “Scientific Acceptability” serves as the secondary threshold criterion. If the measure is endorsed by NQF, the measure must also pass all Scientific Acceptability subcriteria or else NQF will not evaluate it against the remaining criteria when NQF considers the measure for continued endorsement.
- ◆ 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, and some criteria are more accurately evaluated after being in use for some time.
- ◆ 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. At reevaluation, the measure contractor has more information about the strengths and potential weaknesses of a measure than was available at development. The reevaluating contractor is strongly encouraged to take advantage of the situation.
- ◆ 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 maintenance, there are two specific milestones at which a formal measure evaluation report must be submitted to CMS to move the measure development process forward:

1. When recommending disposition of a measure after a comprehensive reevaluation.
2. When recommending disposition of a measure after an ad hoc review

2.5 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 EHR measures are provided in the appendices. 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 objectively 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.

2.6 Procedure

The measures are evaluated to determine the degree to which each measure continues to be 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, feasibility of the measure. In addition, during maintenance phase opportunities for harmonization can be identified including the extent to which the measure being maintained can be harmonized with related measures, or justification of measure differences with competing measures.

Follow the steps below to formally evaluate each measure and update or complete a measure evaluation report.

Recommending disposition of measures at Comprehensive Reevaluation

Refer to the Information Gathering During Maintenance Phase and the Measure Maintenance sections for detailed instructions on updating the MIF and Measure Justification. In the Importance section of the Measure Justification, update the information to demonstrate the business (or economic or social) case for the measure. Use the information from the MIF, Measure Justification, and the analysis of the measure's performance as the primary bases for evaluating and recommending measures for approval.

Step 1: Assess the measure against all five criteria

Using the articles and other data found during the information gathering process, assess the measure against the four criteria to determine if it is still the best available way to determine the quality of care. Questions to consider in this assessment include, but are not limited to:

- ◆ Is the focus of this measure still considered an important issue in the care of these patients?
- ◆ Is there still a gap in care between the measure's performance rates and the highest expected rate, or is there still a significant disparity in care for some patient groups?
- ◆ Is the evidence supporting the measure focus of sufficient quality, quality and consistency of results?
- ◆ To what extent do the performance data and audit results support this measure's reliability and validity?
- ◆ Does the unsolicited feedback (questions or comments about the measure while it has been in use) indicate that the measure needs more precise specifications?
- ◆ Have health outcomes related to this measure improved, and is there a reasonable relationship between that improvement and this measure?
- ◆ Is there a more efficient or effective way to collect the data necessary for this measure?

Step 2: Document the assessment from Step 1

The Measure Evaluation Guidance is available to assist the contractor and the TEP in their evaluation of the measure, if the contractor chooses to use it. The Measure Evaluation report or a similarly formatted document should be completed after evaluating the measure.

Step 3: Present the TEP with the Measure Evaluation Guidance or other documentation for formal evaluation to recommend a disposition for the measure

Provide TEP members with appropriate materials for evaluating the measures before the meeting. Instruct the TEP on the use of the measure evaluation criteria. The Measure Evaluation Guidance may be used to apply the criteria to the measures. Propose to the TEP the potential measures and the rationale behind them as well as any outstanding controversies about the measures. Depending on the specifics of the measure contract, there are several ways the measures can be evaluated:

- ◆ The measure contractor may conduct a preliminary evaluation of the measures using the Measure Evaluation Guidance to complete a draft Measure Evaluation report. These drafts can be presented to the TEP for discussion and input.
- ◆ The measure contractor may send the Measure Evaluation Guidance and instructions on measure evaluation to the TEP members in advance and ask them to evaluate the measures. Before the meeting, the TEP members can submit their assessments. The measure contractor may aggregate the TEP evaluations prior to the meeting and focus the discussion on areas that need further debate.
- ◆ The measures are aggregated during the TEP meeting. This option requires a strong facilitator to focus the discussion.
- ◆ Or, as directed by the COR/GTL.

Step 4: Based on the TEP deliberations, the measure contractor updates the MIF, Measure Justification, formal measure evaluation ratings, and the measure evaluation report for each measure

Step 5: Develop a recommended disposition for review by the COR/GTL

Review, synthesize and incorporate the TEP's input into a measure evaluation report, including the rating for the extent to which the measure meets the evaluation criteria. Submit a measure evaluation report to the COR/GTL for each measure evaluated by the TEP and recommend whether the measure should be continued. Refer to the Comprehensive Reevaluation section for further instructions.

Recommending a disposition for a measure after an ad hoc review

Refer to the Ad Hoc Review section for the standardized process. The following steps address only the sub-process of evaluation.

Step 1: Identify which criteria or subcriteria are directly involved in the ad hoc review. Determine if other criteria or subcriteria are indirectly involved in such a way that they may need to be assessed, as well

If the ad hoc review is triggered by a critique of the measure's validity (ability to address the problem intended), the issue may rest on a lack of reliability (ability to be repeated with similar results), thus making the measure also invalid.

Step 2: Assess the measure's strength relative to the specific criteria or subcriteria identified in Step 1

It is not necessary to address the full range of evaluation criteria. Focus on only those criteria or subcriteria that have triggered the ad hoc review.

Step 3: Update the MIF and Measure Justification for the measure based on the new information triggering the ad hoc review

Step 4: Provide the TEP members the opportunity to evaluate and rate the measures based on the new information

The TEP should be encouraged to remain focused on only those criteria or subcriteria involved in the ad hoc review. This is not an opportunity to reopen a broad discussion of the measure's suitability.

The TEP evaluation can be done in several different ways as explained in Step A3 above. Gather feedback from the TEP and apply the Measure Evaluation Criteria to determine if the measure passes the measure evaluation criteria. The Measure Evaluation Guidance in the appendices of this section can be used or the contractor can use an alternative approach based on the consideration of the COR/GTL.

Step 5: If the measure is under review because of the Scientific Acceptability criterion, identify ways to refine and update the measure that would allow it to pass this evaluation. Consider testing the refined measure

If the measure no longer passes the Scientific Acceptability criterion, identify ways that the measure can be refined as necessary based on the results of the TEP's discussion. Retest the measure if possible (refer to the Measure Testing During Maintenance Phase section for details). Update the MIF and Measure Justification based on the results of the retesting and have the TEP rate the measure again. Continue refining, retesting and updating the measure evaluation to the greatest extent possible. Note that the measure must pass the Scientific Acceptability criterion in order to retain NQF endorsement.

Step 6: If the measure is under review because of the Usability and Feasibility criteria, based on the qualitative and quantitative identify ways to refine and update the measure that would allow it to pass this evaluation. Consider testing the refined measure

Continue to refine and retest the measure to the extent possible based on the TEP's deliberation and, if possible, test results. Update the MIF and Measure Justification based on the results of the new information and have the TEP rate the measure again.

Review, synthesize and incorporate the TEP's input into a measure evaluation report including the rating for the extent to which the measure meets the evaluation criteria. Submit a measure evaluation report to the COR/GTL for each measure evaluated by the TEP and recommend an appropriate disposition for the measure. In order to be recommended for continued use, a measure should pass all of the importance and scientific acceptability criteria.

2.8 Overview of Appendices

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

Appendix 2a: Tools for Individual Measure Evaluations

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 Measure Evaluation Guidance should be used as the guide to determine the ratings for the criteria ratings that are reported in the Measure Evaluation Report, Appendix 2e.

Appendix 2b: Tools for Composite Measures Evaluations

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 Measure Evaluation Guidance document should be used as the guide to determine the ratings for the criteria ratings that are reported in the Measure Evaluation Report, Appendix 2e.



Appendix 2c: Tools for eMeasures Evaluation

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 Tool Guidance—This tool corresponds to the eMeasure 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 Measure Evaluation Guidance should be used as the guide to determine the ratings for the criteria ratings that are reported in the Measure Evaluation Report, Appendix 2e.

Appendix 2d: Tools for Cost and Resource Use Measures Evaluation

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 Measure Evaluation Guidance has not been developed for cost and resource use measures; however, using the evaluation criteria, the Measure Evaluation Report can be used to document evaluation of the measure.

Appendix 2e: 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 2011. Available at: http://www.qualityforum.org/projects/efficiency_resource_use_1.aspx#t=2&s=&p=&e=1 Last accessed: June 2011.

2a Measure Evaluation Criteria and Subcriteria for Individual Measures

Adapted from National Quality Forum Measure Evaluation Criteria⁵

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:

1c1 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 → 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)

⁵National Quality Forum Measure Evaluation Criteria January 2011. Available at: http://www.qualityforum.org/docs/measure_evaluation_criteria.aspx. Accessed August 1, 2012.

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).

1c2 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 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 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 could 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).

2b Measure Evaluation Tool (MET) for Composite Measures

Evaluation criteria adapted from NQF

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 (MER).

Importance—Impact, Opportunity, Evidence

Subcriterion	Pass	Fail
<p>1a. High Impact</p>	<p>The measure focus addresses a specific national health goal/priority identified by one or more of the following:</p> <ul style="list-style-type: none"> ◆ CMS/HHS ◆ Legislative mandate ◆ NQF’s National Priorities Partners <p>OR</p> <p>The measure focus has high impact on health care as demonstrated by one or more of the following:</p> <ul style="list-style-type: none"> ◆ 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 case⁶) ◆ Patient/societal consequences of poor 	<p>The measure does not directly address a national health goal/priority.</p> <p>AND</p> <p>The data do not indicate it is a high impact aspect of health care, or is unknown.</p>

⁶A business case is described the Information Gathering section—Appendix-C.

quality 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)	Performance gap is unknown,
	OR Data demonstrate overall poor performance across providers or population groups (disparities).	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, 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	<i>High does not apply. If this exception is applicable, 1c is either rated Moderate or Low</i>	If there is no empirical evidence, expert opinion is systematically assessed with agreement that the benefits to	For expert opinion evidence that is systematically assessed, lack of agreement that benefits to patients

Empirical Body of Evidence – (structure and process measures)	patients greatly outweigh potential harms.	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 <i>(potential exception)</i>	No empirical evidence <i>(potential exception)</i>	No empirical evidence <i>(potential exception)</i>	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: All of the subcriteria (1a, 1b, 1c) are rated “Pass”.

Fail: Any 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.;	<i>Moderate does not apply. This subcriterion is either rated High or Low.</i>	One or more measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) are <u>ambiguous</u> 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

Subcriterion	High	Moderate	Low
			the score, etc.;
2a2. Reliability testing	<p>Empirical evidence of reliability of <i>BOTH data elements AND measure score within acceptable norms.</i></p> <p>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;</p> <p>AND</p> <p>Measure score: appropriate method, scope, and reliability statistics for score computation or risk adjustment</p>	<p>Empirical evidence of reliability <i>within acceptable norms for either critical data elements OR measure score.</i></p> <p>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;</p> <p>OR</p> <p>Measure score: appropriate method, scope, and reliability statistics for score computation or risk adjustment</p>	<p>Empirical evidence (using appropriate method and scope) of <i>unreliability for either data elements OR measure score, i.e.,</i> statistical results outside of acceptable norms</p> <p>OR</p> <p>Inappropriate method or scope of reliability testing</p>

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	<p>The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1c) under <i>Importance to Measure and Report</i></p>	<p><i>Moderate does not apply. This subcriterion is either rated High or Low</i></p>	<p>The measure specifications <u>do not</u> reflect the evidence cited under <i>Importance to Measure and Report</i> as noted above;</p>
2b2.	Empirical evidence of validity of	Empirical evidence of validity	Empirical evidence (using

Subcriterion	High	Moderate	Low
Validity testing	<p>BOTH data elements AND measure score within acceptable norms:</p> <p>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;</p> <p>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:</p> <ul style="list-style-type: none"> • Construct validity • Discriminative validity/Contrasted groups • Predictive validity • Convergent validity • Reference strategy/Criterion validity 	<p><i>within acceptable norms for either critical data elements OR</i> measure score</p> <p>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;</p> <p>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:</p> <ul style="list-style-type: none"> • Construct validity • Discriminative validity/Contrasted groups • Predictive validity • Convergent validity <p>Reference strategy/Criterion validity</p> <p>OR</p> <p>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,</p>	<p>appropriate method and scope) of invalidity for either data elements OR measure score, i.e., statistical results outside of acceptable norms</p> <p>OR</p> <p>Systematic assessment of face validity of measure resulted in <i>lack of consensus</i> as to whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and poor quality.</p> <p>OR</p> <p>Inappropriate method or scope of validity testing (including inadequate assessment of face validity)</p>

Subcriterion	High	Moderate	Low
		and whether they can be used to distinguish between good and poor quality.	
<p>2b2.</p> <p>Validity testing – threats to validity</p>	<p>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</p>	<p><i>Moderate does not apply. This subcriterion is either rated High or Low</i></p>	<p>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</p> <p>OR</p> <p>Threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or “incorrect” data) are likely and are NOT empirically assessed</p>
<p>2b3.</p> <p>Exceptions</p>	<p>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;</p> <p>AND</p> <p>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);</p> <p>AND</p> <p>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</p>	<p><i>Moderate does not apply. This subcriterion is either rated High or Low</i></p>	<p>Exceptions are not supported by evidence,</p> <p>1. OR</p> <p>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);</p> <p>OR</p> <p>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;</p>

Subcriterion	High	Moderate	Low
	<p>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).</p>		
<p>2b4. Risk Adjustment <i>For outcome measures and other measures (e.g., resource use) when indicated:</i></p>	<p>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;</p> <p>OR</p> <p>rationale/data support not using risk adjustment.</p>	<p>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.</p>	<p>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.</p>
<p>2b5. Meaningful</p>	<p>Data analysis demonstrates that methods for scoring and analysis allow for identification of statistically significant and practically/clinically meaningful differences in performance.</p> <p>OR</p> <p>there is evidence of overall less than optimal performance.</p>	<p>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.</p>	<p>Data analysis was not conducted to identify statistically significant and practically/clinically meaningful differences in performance,</p> <p>OR</p> <p>Methods for scoring and analysis are not appropriate to identify such difference.</p>
<p>2b6. Comparable results</p>	<p>If multiple data sources/methods are allowed (specified in the measure), data and analysis demonstrate that they produce comparable results</p>	<p>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</p>	<p>Multiple data sources/methods are allowed and it is unknown if they produce comparable results,</p> <p>OR</p> <p>testing demonstrates they do not produce comparable results</p>

2c. Disparities

High	Moderate	Low
<p>Data indicate disparities do not exist;</p> <p>2. OR</p> <p>if disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results</p>	<p>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</p>	<p>Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities;</p> <p>AND</p> <p>rationale/data are not provided to justify why stratification is neither necessary nor feasible</p>

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
<p>3a.</p> <p>Public Reporting</p>	<p>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).</p>	<p>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)</p> <p>OR</p> <p><i>When measure is being rated during its initial development:</i></p> <p>A rationale for how the measure performance results will be meaningful, understandable, and useful for public reporting.</p>	<p>The measure is not in use and has not been tested for usability;</p> <p>OR</p> <p>Testing demonstrates information produced by the measure is not meaningful, understandable, and useful for public reporting</p> <p>OR</p> <p><i>When measure is being rated during its initial development:</i></p> <p>A rationale for how the measure performance results will be meaningful, understandable, and useful for public reporting is not provided.</p>
<p>3b.</p> <p>Quality Improvement</p>	<p>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).</p>	<p>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).</p> <p>OR</p> <p><i>When measure is being rated during its initial development:</i></p> <p>A rationale for how the measure performance results will be meaningful, understandable, and</p>	<p>The measure is not in use and has not been tested for usability;</p> <p>OR</p> <p>Testing demonstrates information produced by the measure is not meaningful, understandable, and useful for public reporting</p> <p>OR</p> <p><i>When measure is being rated during its initial development:</i></p> <p>A rationale for how the measure performance results will be</p>

useful for quality improvement.

meaningful, understandable, and useful for quality improvement is **not provided**.

Guidelines for Summary Rating: Usability

Summary Rating: Usability

Pass: The measure rates “Moderate” 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 (<i>clinical measures only</i>)	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 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)



2b.1 Measure Evaluation Criteria and Subcriteria for Composite Measures

Adapted from National Quality Forum Measure Evaluation Criteria⁷ and Composite Measure Evaluation Framework⁸

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.

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.

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

⁸National 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.

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: EHR measures 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 could 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:

<p>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).</p>	
<p>5a. Related measure: The measure specifications for this measure are completely harmonized with a related measure.</p>	
<p>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)</p>	

2b.2 Measure Evaluation Tool (MET) for Composite Measures

(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 (MER).

Importance—Impact, Opportunity, Evidence

Subcriterion	Pass	Fail
<p>1a. High Impact</p>	<p>The measure focus addresses a specific national health goal/priority identified by one or more of the following:</p> <ul style="list-style-type: none"> CMS/HHS Legislative mandate NQF’s National Priorities Partners <p>OR</p> <p>The measure focus has high impact on health care as demonstrated by one or more of the following:</p> <ul style="list-style-type: none"> 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 case⁹) Patient/societal consequences of poor quality regardless of cost (social case) 	<p>The measure does not directly address a national health goal/priority.</p> <p>AND</p> <p>The data do not indicate it is a high impact aspect of health care, or is unknown.</p>

⁹A business case is described the Information Gathering section.

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

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.	<p>Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect;</p> <p>OR</p> <p>RCTs without serious flaws that introduce bias, but with either indirect evidence, or imprecise estimate of effect.</p>	<p>RCTs with flaws that introduce bias</p> <p>OR</p> <p>Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations</p>
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.	<p>Estimates of benefits and harms to patients are consistent in direction, but differ in magnitude across studies in the body of evidence;</p> <p>OR</p>	<p>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;</p> <p>OR</p> <p>For expert opinion evidence that is</p>

Subcriterion	High	Moderate	Low
		<p>If only one study, the estimate of benefits greatly outweighs the estimate of potential harms,</p> <p>OR</p> <p>For expert opinion that is systematically assessed, agreement that benefits to patients clearly outweigh potential harms.</p>	<p>systematically assessed, lack of agreement that benefits to patients clearly outweigh potential harms.</p>
1c: Potential Exception to Empirical Body of Evidence – (structure and process measures)	<i>High does not apply. If this exception is applicable, 1c is either rated Moderate or Low</i>	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 <i>(potential exception)</i>	No empirical evidence <i>(potential exception)</i>	No empirical evidence <i>(potential exception)</i>	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	The component measures are determined to meet	a component measure is not important enough in

Component ratings	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 important component of the composite.	its own right as an individual measure, And is not an 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: All of the subcriteria (1a, 1b, 1c, 1d, and 1e) are rated “Pass”.

Fail: Any 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
--------------	------	----------	-----

Subcriterion	High	Moderate	Low
<p>2a1.</p> <p>Specifications well defined and precisely specified</p>	<p>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.;</p> <p>AND</p> <p>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</p>	<p>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.;</p> <p>AND</p> <p>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.</p>	<p>One or more measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) are <u>ambiguous</u> 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.;</p> <p>OR</p> <p>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</p> <p>OR</p> <p>Analysis of missing component scores does not support the specifications for scoring/aggregation and handling of missing component scores.</p> <p>OR</p> <p>Analysis not performed</p>
<p>2a2.</p> <p>Reliability testing</p>	<p>Empirical evidence of reliability of <u>BOTH data elements AND measure score</u> within acceptable norms.</p> <p><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.,</p>	<p>Empirical evidence of reliability <u>within acceptable norms</u> for either critical data elements OR measure score.</p> <p><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.,</p>	<p>Empirical evidence (using appropriate method and scope) of <u>unreliability</u> for either data elements OR measure score, i.e., statistical results outside of acceptable norms</p> <p>OR</p> <p>Inappropriate method or scope of reliability testing</p>

Subcriterion	High	Moderate	Low
	gender, age, date of admission); OR <i>may forego data element reliability testing if data element validity was demonstrated;</i>	gender, age, date of admission); OR <i>may forego data element reliability testing if data element validity was demonstrated;</i>	
	AND <u>Measure score:</u> appropriate method, scope, and reliability statistics for score computation or risk adjustment	OR <u>Measure score:</u> 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 <i>Importance to Measure and Report</i>	<i>Moderate does not apply. This subcriterion is either rated High or Low</i>	The measure specifications <u>do not</u> reflect the evidence cited under <i>Importance to Measure and Report</i> as noted above;
2b2. Validity testing	Empirical evidence of validity of <u>BOTH data elements AND measure score within acceptable norms:</u> <u>Data element:</u> appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; <u>Measure score:</u> Evidence that	Empirical evidence of validity <u>within acceptable norms</u> for <u>either critical data elements OR measure score</u> <u>Data element:</u> appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; <u>Measure score:</u> Evidence that	Empirical evidence (using appropriate method and scope) of <u>invalidity</u> for <u>either data elements OR measure score</u> , i.e., statistical results outside of acceptable norms OR Systematic assessment of face validity of measure resulted in <u>lack of consensus</u> as to whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and

Subcriterion	High	Moderate	Low
	<p>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:</p> <p>Construct validity</p> <p>Discriminative validity/Contrasted groups</p> <p>Predictive validity</p> <p>Convergent validity</p> <p>Reference strategy/Criterion validity</p>	<p>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:</p> <p>Construct validity</p> <p>Discriminative validity/Contrasted groups</p> <p>Predictive validity</p> <p>Convergent validity</p> <p>Reference strategy/Criterion validity</p> <p>OR</p> <p>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.</p>	<p>poor quality.</p> <p>OR</p> <p>Inappropriate method or scope of validity testing (including inadequate assessment of face validity)</p>
<p>2b2.</p> <p>Validity testing – threats to validity</p>	<p>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</p>	<p><i>Moderate does not apply. This subcriterion is either rated High or Low</i></p>	<p>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</p>

Subcriterion	High	Moderate	Low
	<p>results are not biased</p>		<p>OR</p> <p>Threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or “incorrect” data) are likely and are NOT empirically assessed</p>
<p>2b3. Exceptions</p>	<p>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;</p> <p>AND</p> <p>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);</p> <p>AND</p> <p>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).</p>	<p><i>Moderate does not apply. This subcriterion is either rated High or Low</i></p>	<p>Exceptions are not supported by evidence,</p> <p>OR</p> <p>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);</p> <p>OR</p> <p>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;</p>
<p>2b4. Risk Adjustment</p>	<p>An evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is</p>	<p>An evidence-based risk-adjustment strategy is specified consistent with the evaluation</p>	<p>A risk-adjustment strategy is not specified AND would be absolutely necessary for the</p>

Subcriterion	High	Moderate	Low
For outcome measures and other measures (e.g., resource use) when indicated:	<p>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;</p> <p>OR</p> <p>rationale/data support not using risk adjustment.</p>	<p>criteria, HOWEVER it uses a method that is less than ideal, but acceptable.</p>	<p>measure to be fair and support valid conclusions about the quality of care.</p>
2b5. Meaningful	<p>Data analysis demonstrates that methods for scoring and analysis allow for identification of statistically significant and practically/clinically meaningful differences in performance.</p> <p>OR</p> <p>there is evidence of overall less than optimal performance.</p>	<p>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.</p>	<p>Data analysis was not conducted to identify statistically significant and practically/clinically meaningful differences in performance,</p> <p>OR</p> <p>Methods for scoring and analysis are not appropriate to identify such difference.</p>
2b6. Comparable results	<p>If multiple data sources/methods are allowed (specified in the measure), data and analysis demonstrate that they produce comparable results</p>	<p>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</p>	<p>Multiple data sources/methods are allowed and it is unknown if they produce comparable results,</p> <p>OR</p> <p>testing demonstrates they do not produce comparable results</p>
2b7. Components fit the conceptual construct;	<p>Component item/measure analysis demonstrates that the included component items/measures fit the conceptual construct;</p> <p>OR</p> <p>justification and results for alternative analyses are provided.</p>	<p>Component item/measure analysis demonstrates that the included component items/measures are somewhat related to conceptual construct;</p>	<p>Component item/measure analysis does not demonstrates that the included component items/measures fit the conceptual construct;</p> <p>OR</p> <p>justification and results for alternative analyses is not provided.</p>

Subcriterion	High	Moderate	Low
2b8. Components contribute to the variation	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	Data analysis was conducted but did not demonstrate that the included components contribute to the variation in the overall composite score; HOWEVER the methods for scoring and analysis are appropriate to identify such differences if they exist	Data analysis was not conducted to demonstrate that the included components contribute to the variation in the overall composite 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,	Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities; HOWEVER the rationale/data	Disparities in care have been identified, but measure specifications, scoring, and analysis do not allow for identification of disparities;

High	Moderate	Low
scoring, and analysis allow for identification of disparities through stratification of results	justify why stratification is neither necessary nor feasible	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
<p>3a</p> <p>Byproduct of care (<i>clinical measures only</i>)</p>	<p>The required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery (e.g., BP reading, diagnosis).</p>	<p>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.</p>	<p>The required data are not generated during care delivery and are difficult to collect or require special surveys or protocols.</p>
<p>3b</p> <p>Electronic data</p>	<p>The required data elements are available in electronic health records or other electronic sources.</p>	<p>If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.</p>	<p>The required data elements are not available in electronic sources</p> <p>AND</p> <p>There is no credible, near-term path to electronic collection specified.</p>
<p>3c</p> <p>Inaccuracies, errors, or unintended consequences</p>	<p>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</p>	<p>There is moderate susceptibility to inaccuracies, errors, or unintended consequences,</p> <p>AND/OR</p> <p>They are more difficult to detect through auditing.</p>	<p>Inaccuracies, errors, or unintended consequences have been demonstrated,</p> <p>AND</p> <p>They are not easily detected through auditing.</p>
<p>3d</p> <p>Data collection strategy</p>	<p>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.</p>	<p>The measure is not in operational use; HOWEVER testing demonstrates the data collection strategy can be implemented with minimal difficulty or additional resources.</p>	<p>The measure is not in operational use,</p> <p>AND</p> <p>Testing indicates the data collection strategy was difficult to implement and/or requires substantial additional resources.</p>

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	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),	Measure is not currently used in at least one accountability application and performance improvement; HOWEVER there is a credible plan for implementation provided.	Credible plan for implementation or rationale for potential use of performance results in quality improvement is not provided .
And Improvement	AND Performance improvement	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 Description of how benefits outweigh them were not provided OR Description of actions taken to mitigate them were not provided
And Benefits of Performance Measure	OR No unintended consequences were identified during testing No evidence of unintended consequences to individuals and	AND Unintended consequences to individuals or populations were identified AND Description of how benefits	

Subcriterion	High	Moderate	Low
	populations have been reported since implementation.	outweigh them were provided or description of actions taken to mitigate them were provided	
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.	<i>Moderate does not apply. This subcriterion is either rated High or Low</i>	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 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 “Moderate” 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)

2c.1 Measure Evaluation Criteria and Subcriteria for New eMeasures Specified for EHRs and Endorsed Measures Re-specified for Use With EHR

Adapted from National Quality Forum Measure Evaluation Criteria¹⁰

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

1. 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.

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

¹⁰National 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:

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: EHR measures 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

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

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

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 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.

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

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 could 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)

2c.2 Measure Evaluation Tool (MET) for eMeasures

(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 (MER).

Importance—Impact, Opportunity, Evidence

Subcriterion	Pass	Fail
1a. High Impact	<p>The measure focus addresses a specific national health goal/priority identified by one or more of the following:</p> <ul style="list-style-type: none"> ◆ CMS/HHS ◆ Legislative mandate ◆ NQF’s National Priorities Partners <p>OR</p> <p>The measure focus has high impact on health care as demonstrated by one or more of the following:</p> <ul style="list-style-type: none"> ◆ 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 case¹¹) ◆ Patient/societal consequences of poor quality regardless of cost (social case) 	<p>The measure does not directly address a national health goal/priority.</p> <p>AND</p> <p>The data do not indicate it is a high impact aspect of health care, or is unknown.</p>
1b. Performance Gap	<p>Evidence exists to substantiate a quality problem and opportunity for improvement (i.e., data demonstrate considerable variation)</p> <p>OR</p> <p>Data demonstrate overall poor performance across providers or population groups (disparities).</p>	<p>Performance gap is unknown,</p> <p>OR</p> <p>There is limited or no room for improvement (no variability across providers or population groups and overall good performance).</p>

¹¹A business case is described the Information Gathering section—Appendix 5-C.

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, 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)	<i>High does not apply. If this exception is applicable, 1c is either rated Moderate or Low</i>	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 evidence links the health outcome to a known	A rationale supports the relationship of the health	No rationale is given that supports the relationship of the health

Subcriterion	Pass	Fail
Empirical Body of Evidence – (health outcome measures)	process or structure of care.	outcome to at least one healthcare structure, process, intervention, or service.
		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 <i>(potential exception)</i>	No empirical evidence <i>(potential exception)</i>	No empirical evidence <i>(potential exception)</i>	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: All of the subcriteria (1a, 1b, 1c) are rated “Pass”.

Fail: Any 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
<p>2a1. Specifications well defined and precisely specified</p>	<p>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.</p>	<p><i>Moderate does not apply. This subcriterion is either rated High or Low</i></p>	<p>One or more EHR measure specifications are ambiguous or do not use data elements from the QDM.</p> <p>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.</p>
<p>2a2. Reliability testing</p>	<p>Empirical evidence of reliability of <u>both data element AND measure score</u> within acceptable norms:</p> <p><u>Data element</u>: reliability (repeatability) assured with computer programming—must test data element validity AND <u>Measure score</u>: appropriate method, scope, and reliability statistic within acceptable norms</p>	<p>Empirical evidence of reliability <u>within acceptable norms</u> for <u>either data elements OR measure score</u>.</p> <p><u>Data element</u>: reliability (repeatability) assured with computer programming—must test data element validity AND <u>Measure score</u>: appropriate method, scope, and reliability statistic within acceptable norms</p>	<p>Empirical evidence of <u>unreliability</u> for <u>either data elements OR measure score</u>—i.e., statistical results outside of acceptable norms</p>

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 <i>Importance</i>	<i>Moderate does not apply. This subcriterion is either rated High or Low</i>	The measure specifications <u>do not</u> reflect the evidence cited under <i>Importance to Measure and Report</i> as noted above;
2b2. Validity testing	<p>Empirical evidence of validity of <u>both data elements AND measure score</u> within acceptable norms:</p> <p><u>Data element:</u> validity demonstrated by analysis of agreement between data elements electronically extracted and data elements visually abstracted from the <u>entire</u> 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</p> <p><u>Measure score:</u> appropriate method, scope, and validity testing result within acceptable norms</p>	<p><u>Empirical evidence</u> of validity within acceptable norms for <u>either data elements OR measure score</u></p> <p><u>Data element:</u> validity demonstrated by analysis of agreement between data elements electronically extracted and data elements visually abstracted from the <u>entire</u> 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</p> <p><u>Measure score:</u> appropriate method, scope, and validity testing result within acceptable norms</p> <p>OR <u>Systematic assessment of face validity of measure score as a quality indicator explicitly addressed and found <u>substantial agreement</u> 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.</u></p>	<p>Empirical evidence (using appropriate method and scope) of <u>invalidity</u> for <u>either data elements OR measure score</u>, i.e., statistical results outside of acceptable norms</p> <p>OR Systematic assessment of face validity of measure resulted in <u>lack of consensus</u> as to whether measure scores provide an accurate reflection of quality, and whether they can be used to distinguish between good and poor quality.</p> <p>OR Inappropriate method or scope of validity testing (including inadequate assessment of face validity)</p>

Subcriterion	High	Moderate	Low
<p>2b2. Validity testing – threats to validity</p>	<p>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.</p>	<p><i>Moderate does not apply. This subcriterion is either rated High or Low</i></p>	<p>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</p>
<p>2b3. Exceptions</p>	<p>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).</p>	<p><i>Moderate does not apply. This subcriterion is either rated High or Low</i></p>	<p>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;</p>
<p>2b4. Risk Adjustment For outcome measures and other</p>	<p>An evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the</p>	<p>An evidence-based risk-adjustment strategy is specified consistent with the evaluation criteria, HOWEVER it uses a method that is less than ideal, but</p>	<p>A risk-adjustment strategy is not specified AND would be absolutely necessary for the measure to be fair and support valid conclusions about the</p>

Subcriterion	High	Moderate	Low
<i>measures (e.g., resource use) when indicated:</i>	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.	acceptable.	quality of care.
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 <u>do not</u> use only data elements from the QDM; OR Crosswalk of the EHR measure specifications as noted above identifies that they <u>do not</u> 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
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
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 <i>When measure is being rated during its initial development:</i> 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 <i>When measure is being rated during its initial development:</i> 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 <i>When measure is being rated during its initial development:</i> 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 <i>When measure is being rated during its initial development:</i> A rationale 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 “Moderate” 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 (<i>clinical measures only</i>)	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

2d Measure Evaluation Criteria and Subcriteria for Resource Use Measures

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 included must be conceptually coherent and consistent with the purpose.

¹²The 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:

(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 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

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

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-than-optimal performance may not demonstrate much variability across providers.

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

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 could 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.

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

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)

2e Measure Evaluation Report

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/Exceptipn		
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:

3. Harmonization During Maintenance Phase

3.1 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 parts 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 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.²

The term “measure alignment” is also used to describe harmonization and related activities. Alignment of measures across programs and agencies can 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 standardizing the measure specifications (numerator and denominator inclusions and exclusions) across programs, standardizing other aspects of measures

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

²Ibid.

and measure sets such as the value sets from which measures are constructed, measurement periods used across programs, and streamlining reporting methods.

3.2 Procedure

When reevaluating and updating 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 is to ensure that the measure being updated is harmonized with other existing measures. Measure contractors should also consider new measures that are currently under development. It should not be assumed that an endorsed measure or other measure that is in use by CMS is better than a new measure. Measure contractors should also consider these new measures during maintenance.

Measures should be harmonized with other existing measures unless there is a compelling reason for not doing so. Harmonization means standardizing 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 ways to measure those concepts.

Harmonizing measure specifications during measure development is more efficient than after a measure has been fully developed and specified. However, the measure maintenance processes allow opportunities to harmonize measures that are already fully developed and in use.

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

Step 1: Decide whether harmonization is needed

Conduct environmental scan for similar measures already in existence and for related measures. Also, look for measures in development that are similar or related. 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.

Table 9-1 summarizes issues and possible actions to be considered when evaluating measures identified during the environmental scan.³

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

Table 9-1—Harmonization Decisions

	Harmonization Issue	Action
Numerator: Same measure focus Denominator: Same target population	Competing measures	<ul style="list-style-type: none"> ◆ 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	<ul style="list-style-type: none"> ◆ Harmonize on measure focus (Adapted), or ◆ Justify differences, or ◆ Adapt existing measure by expanding the target population.
Numerator: Different measure focus Denominator: Same target population	Related measures	<ul style="list-style-type: none"> ◆ Harmonize on target population, or ◆ Justify differences.
Numerator: Different measure focus Denominator: Different target population	New measures	<ul style="list-style-type: none"> ◆ Develop measure.

Step 2: Update measure specifications

When evaluating the measure specifications, consider various aspects of the measure for potential harmonization. Harmonization often requires close inspection of the detail specification 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 different 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 a 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 the reason and include any literature used to support this decision. Reasons for not harmonizing may include:

- ◆ The science behind the new measure does not support using the same variable(s) found in the existing measure.
- ◆ CMS' 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 the 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, 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 code sets to build a new list.

Step 3: Test scientific acceptability of measure properties

If changes to the measure specifications have been made as a result of harmonization, testing is usually necessary. For further details, please refer to Measure Testing During Maintenance Phase section.

Step 4: NQF maintenance submission

As part of each maintenance review, NQF will evaluate the measure for harmonization potential. NQF projects include both new measure evaluation and maintenance evaluation for measures of related topics. There may be instances where the measure contractor may be unaware of newly developed similar or related measures until they have been submitted to NQF for review. If similar or related measures are identified by NQF, and harmonization has not taken place, or reasons for not doing so adequately justified, 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 continued endorsement.

Step 5: Alignment and harmonization outcomes

CMS has adopted a set of criteria when considering measures, designed to ensure a consistent approach. One of the core criteria for selection and ongoing use of a measure is that the measure promotes alignment with specific program attributes and across CMS and HHS. It is preferable to use the same measure or a measure that has been harmonized in CMS and HHS programs. In some cases, it may be necessary to replace a measure in use for a program, with a similar one that is in more widespread use across the agency or one that is harmonized with measures in other programs.

Alignment should be considered during comprehensive reevaluation and may result in the “Replace” outcome if a more appropriate measure is found. Harmonization with an existing measure during comprehensive reevaluation may result in the “Revise” outcome.

4. Measure Production and Monitoring

4.1 Introduction

The Measure Production and Monitoring section provides a high-level overview of the ongoing tasks necessary to use the measure over time. Where applicable, the reader is directed to other sections for more detailed instructions. The process of production and monitoring measures varies significantly from one measure set to another depending on a number of factors, which may include, but is not limited to:

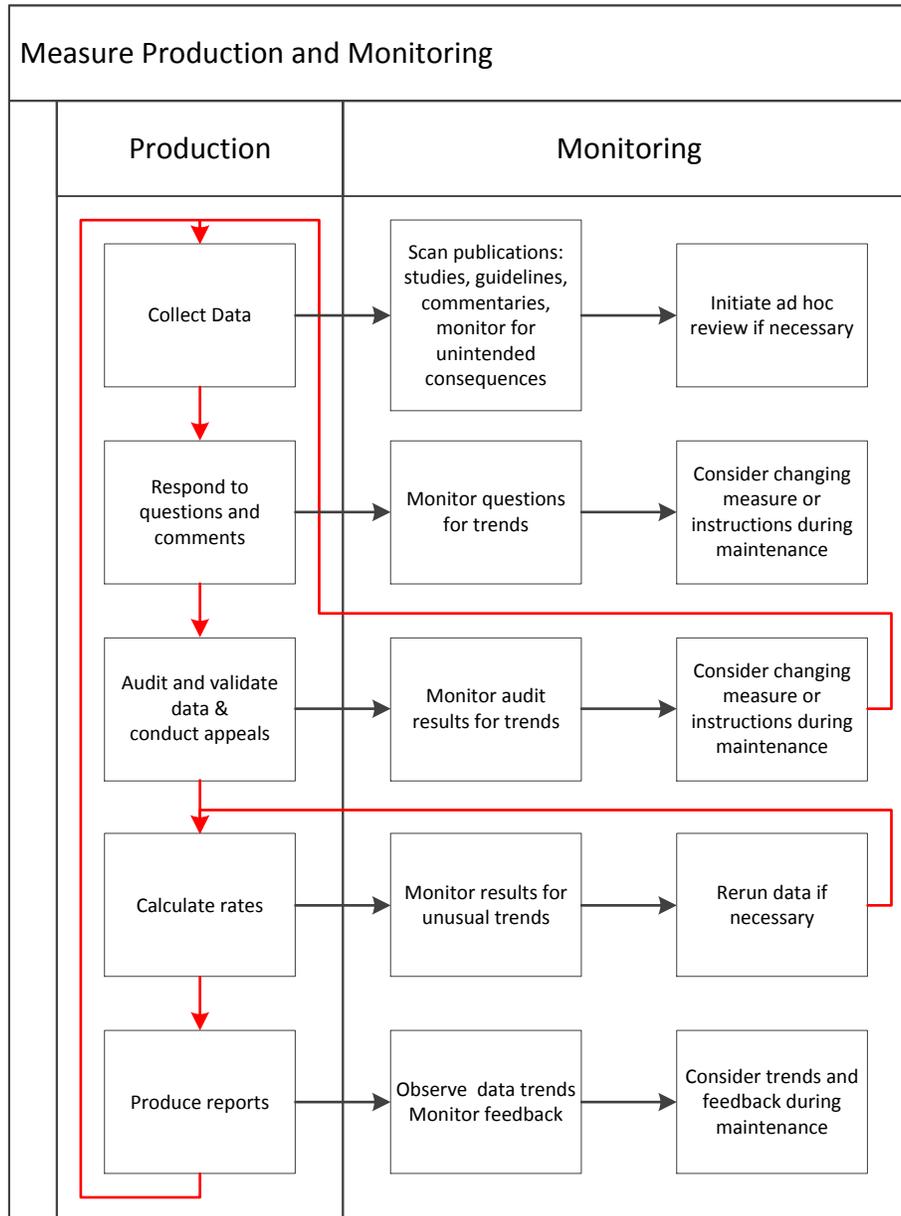
- ◆ Scope of measure implementation.
- ◆ Health care provider being measured.
- ◆ Data collection processes.
- ◆ Ultimate use of the measure (e.g., quality improvement, public reporting, pay-for-reporting, or value-based purchasing).
- ◆ Program in which the measure is used.

While the details of monitoring measures may vary, the high-level processes generally include certain tasks. The intensity or amount of effort involved in each of these tasks may vary and be affected by the factors listed above.

Figure 4-1 represents the high-level process of production and monitoring a measure approved by the Contracting Officer Representative/Government Task Leader (COR/GTL) and implemented in a Centers for Medicare & Medicaid Services (CMS) program.

The work conducted in this section, as with all other sections, will comply with the requirements of the Data Quality Act (DQA) as well as with the Department of Health and Human Services' (HHS) Guidelines for Ensuring the Quality of Information Disseminated to the Public (<http://www.aspe.hhs.gov/infoquality/Guidelines/CMS-9-20.shtml>).

Figure 4-1 Overview of Measure Monitoring Process



4.2 Deliverables

- ◆ Audit and validation reports
- ◆ Audit and validation appeals reports
- ◆ Preview reports, if required by the CMS program using the measure
- ◆ Periodic measure reports
- ◆ Analysis of the measure results
- ◆ Ad hoc reviews, as requested by CMS
- ◆ Periodic environmental scans

- ◆ Program and initiative assessment

4.3 Procedure

Step 1: Conduct data collection and ongoing surveillance

Once the dry run is complete, and any problems that surfaced are resolved, the measure will be fully implemented. This means that the data are collected, calculated, and publicly reported.

As the measure is being used, new studies may be published that address the soundness of the measure. Ongoing surveillance of the literature should be conducted throughout the implementation process. These studies should be gathered and reviewed. If a study suggests a significant concern about the evidence on which the measure is based, determine if an ad hoc review is necessary. Refer to the Ad Hoc Review section for a description of the standardized procedure.

Pay particular attention to any organizations that issue clinical guidelines that are relevant to the measure. If the measure is based on a particular set of guidelines, the guideline writers should be closely monitored for any indication that they are planning on making changes to the guidelines. If the measure is not based on guidelines, the focus should be on any scientific or clinical organization that might issue such guidelines or other statements regarding the scientific basis of the measure. These guideline changes or other statements may necessitate an ad hoc review.

In addition to publications in medical and scientific publications, the general media should be scanned for articles and commentaries about the measure.

After data collection begins, monitor for unintended consequences. Look for articles or studies describe in unintended consequences in literature and identify any unusual trends in data suggesting unintended consequences. If significant unintended consequences are identified, an ad hoc review may be necessary, especially if patient safety is the concern.

Step 2: Respond to questions about the measure

The measure contractor monitoring the measure is responsible for reviewing any stakeholder feedback and responding to it in a timely manner. This stakeholder feedback may include questions or comments about the measure or the program in which the measure is being used. This feedback may come through various systems such as RightNow, or regular email. Assuming the submitter has provided contact information, the measure contractor receiving the feedback should reply immediately letting the submitter know that the feedback has been received and is being reviewed. Within two weeks of the submission date, the measure contractor should provide either a final response to the submitter or a status update to let the submitter know what is happening regarding the feedback. All responses will be reviewed by the COR/GTL unless the COR/GTL makes other arrangements. As with the other components of the environmental scan, stakeholder feedback may necessitate an ad hoc review. Comments may also come in as part of the federal rule making process.

Step 3: Conduct audit and validation and appeals

The auditing and validation processes culminate in a periodic (e.g., monthly, quarterly, or annual) report showing results by provider, aggregated by area, state, or region (as determined by the COR/GTL), and national rates. These reports may be used to document the validity of the provider's result and may be included as criteria in accountability (public reporting or value-based purchasing) programs.

Careful assessment of the audit and validation results may reveal patterns that may require adjustment of the technical specifications. If the measure is to be used for accountability programs (public reporting or value-based purchasing), the providers will have an opportunity to challenge audit reports of low validity or reliability. A threshold may be set (e.g., only providers with reliability less than 80 percent) or other criteria may be required before a provider can appeal the audit results.

For hospital measures, most measures can be appealed to the state's Medicare Quality Improvement Organization (QIO), and then to the Provider Reimbursement Review Board (PRRB). For nursing home and home health agency measures, the providers can appeal the measure results to the state survey agencies (SSAs).

Step 4: Produce preliminary reports

For public reporting programs, the results will be released to the providers before they are released to the public. The providers will be allowed a period of time (usually 30 days) to review and respond to the measure results.

The preliminary reports should be monitored for unusual trends. These trends should be investigated. If an error in calculation, the reports should be rerun. If the unexpected results not due to error, the cause should be investigated and reported to CMS. If necessary, CMS has the option of suppressing some or all of the data from appearing on the Web site for a given reporting period (e.g., quarter or year). Data suppression might be necessary due to known problems with a given measure or measure set, suspension of a particular measure (refer to the Measure Maintenance section for a discussion of measure suspension), or data collection issues with a particular provider or group of providers. The decision to suppress data may apply to:

- ◆ All measures in a given measure set (or sets)
- ◆ A particular measure (or measures)
- ◆ A group of providers (e.g., a state or a region)
- ◆ A particular provider (or providers)

Step 5: Report measure results

Once the measure results are calculated and the providers have reviewed them (for public reporting or value-based purchasing programs), the results are released. Depending on the particular program, the reporting process will vary. For quality improvement programs, the results will be released to the providers, possibly with other providers results indicated. The COR/GTL will determine if the other provider results are to be reported anonymously or not. The process by which the information is to be

shared with providers and others, the format of the reports, and support for questions from providers should be established in the rollout plan before implementation occurs.

If the measure results are to be posted on the appropriate Web site, an announcement of the updated site may be made. The display of the measure results on the Web site may require collaboration with quality alliances and will require consumer testing. Other considerations include compliance with Section 508 of the Rehabilitation Act, which requires federal agencies' electronic information to be accessible to people with disabilities.

For value-based purchasing programs, the results will be shared with the appropriate areas within CMS responsible for calculating provider payments in addition to any requirements for public reporting of the data.

Step 6: Monitor and analyze the measure rates and audit findings

The measure performance rates and audit findings will be monitored and analyzed periodically and at least once a year for the following:

- ◆ Overall performance trends
- ◆ Variations in performance, gaps in care, and 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
- ◆ Correlation of the performance data to either confirm the measure's efficacy or identify weaknesses in the measure

Step 7: Perform measure maintenance or ad hoc review, when appropriate

Each measure is reviewed at least annually to ensure that the codes used to identify the populations (denominator, numerator, and exclusions) are current, and to address other minor changes that may be needed. Refer to the Measure Update section for the standardized process.

Each measure is fully reevaluated every three years to ensure that it still meets the measure evaluation criteria. Refer to the Comprehensive Reevaluation section for the standardized process.

As mentioned above, a situation may arise in which a measure must be reviewed before the scheduled measure update or comprehensive reevaluation. In this case, the ad hoc review is conducted. Refer to the Ad Hoc Review section for the standardized procedure, including the process for determining when an ad hoc review is necessary. An ad hoc review may also be initiated by NQF, if the measure is endorsed. The outcome of the ad hoc review will be incorporated into the monitoring cycle at the appropriate place, based on the decision approved by the COR/GTL.

Depending on the outcome of the reevaluation (i.e., retire, retain, revise, suspend, or remove), CMS will make decisions on the continued use of a particular measure in a particular program.

If the National Quality Forum (NQF) has endorsed the measure, the results of the maintenance will be reported to the NQF to reevaluate its endorsement at the time of the NQF maintenance review. The outcome of the NQF review may impact CMS's decision on the continued use of a particular measure in a program.

Step 8: Provide information that CMS can use in measure priorities planning

Lessons learned from the measure rollout, the environmental scan, and ongoing monitoring of the measure should be conveyed to CMS staff. Refer to the Measure Priorities Planning section. In addition, 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 programs or initiatives 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. Performance trends of the measure can be used by the NQF Measures Application Partnership (MAP) to evaluate the use of the same or similar measure in other settings or programs. This evaluation may be done as part of the pre-rulemaking process for the Measures Under Consideration.

5. Measure Maintenance

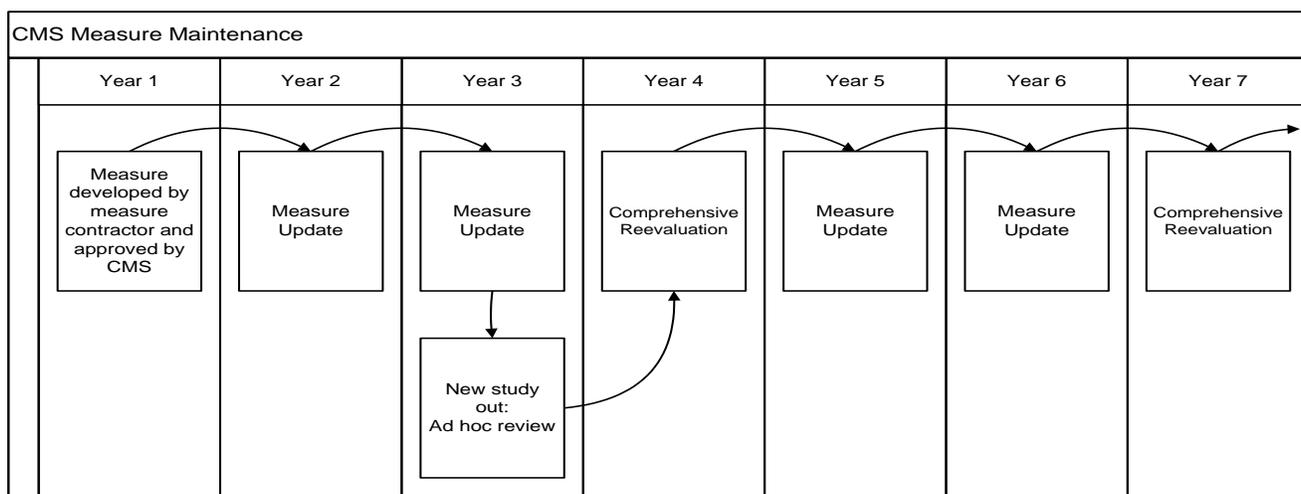
5.1 Introduction

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 (measure development or measure maintenance) and whether the measure should continue to be used. As described in the Measure Production and Monitoring section, ongoing evaluation is part of the implementation process. The information gathered during the monitoring processes will be used in the measure maintenance reviews. The Centers for Medicare & Medicaid Services (CMS) has identified three distinct types of measure maintenance varying primarily in the scope of the review:

- ◆ **Measure Update**—A limited review of the precision of the measure’s specifications focused on ensuring that the procedure, diagnostic, and other codes (CPT, ICD-10-CM, LOINC, etc.) used within the measure are updated as the coding sets themselves are updated. This measure update may be done annually or semiannually. Feedback received regarding the measure’s specifications, reliability, and validity—including the reliability and validity of the measure’s constituent data elements—should be reviewed and may also be addressed at this time.
- ◆ **Comprehensive Reevaluation**—A thorough review of the measure, usually done every three years. A comprehensive review of the literature, measure performance rates, and all feedback received regarding the measure is conducted at this time.
- ◆ **Ad Hoc Review**—A limited review of the measure based on new and urgent information evidence that may have a significant, adverse effect on the measure or its implementation.

Figure 5-1 illustrates how these processes work together over the lifecycle of a measure.

Figure 5-1—CMS Measure Maintenance



This section provides a high-level view of the processes of measure maintenance. In most cases, the reader is directed to the appropriate sections of this Blueprint for more detailed instructions.

5.2 Possible Outcomes

- ◆ **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.

5.3 Measures Owned by Others

Some measures used by CMS are developed and maintained by other organizations that have agreed—sometimes by contract and sometimes with an informal agreement—to allow CMS to use that organization’s measures. These are measures adopted or adapted by CMS for use in certain quality initiatives. One example includes the many measures developed and maintained by the American Medical Association-Physician Consortium for Performance Improvement (AMA-PCPI) for use by CMS. CMS anticipates that most of these measures maintained by measure stewards (i.e., organizations that are not contracted with CMS, but own or maintain measures that are used by CMS) will be maintained in a manner substantially identical to the CMS process. Maintaining measures in a “substantially identical” manner means that the measure steward will:

- ◆ Update any codes that are used in the measure at least annually to account for changes in the coding sets.
- ◆ Periodically conduct a review of the measure’s scientific acceptability, importance, continued feasibility, and usefulness to a stakeholder community. Such a review should be conducted no less frequently than every three years.
- ◆ Have a process for conducting timely reviews if the measure is challenged based on the measure’s scientific acceptability.

CMS reserves the right to review each reevaluated measure to determine its continued suitability for inclusion in the CMS measure set. CMS is not obligated to continue use of a measure it adopts from another measure developer. Because CMS is not the measure owner (steward), it cannot formally

retire a measure. Measures owned by others that are no longer suitable for the CMS program are removed from the program. 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. Before contacting the measure owner, measure contractors should discuss the situation with their COR/GTL.

If CMS decides it wants to make changes to a measure that is currently used in a CMS program and is owned by another organization, consider the following issues:

- ◆ Will the measure owner agree to the changes in the measure specifications that will meet the needs of the current project?
- ◆ If the existing measure is NQF-endorsed, are the changes to the measure significant enough for NQF to initiate an ad hoc review?

Measure performance information may be available from CMS or one of its contractors (either in addition to the data collected by the measure steward or in lieu of such data), in which case this measure performance information will be forwarded to the measure steward at the time of its measure reevaluation.

For a variety of reasons, CMS may use a measure with an owner that has no relationship with CMS. As directed by CMS, the Measures Manager may monitor the measure in question by periodically reviewing the NQF Web site for indications that NQF has been notified of a change to the measure. CMS may also direct the Measures Manager to alert CMS when the measure owner makes changes to the measures.

If a measure used by CMS has no measure owner or measure contractor, CMS will decide whether resources can be allocated to conduct measure maintenance.

5.4 Criteria for Selection, Retirement, Removal and Suspension

In 2012, CMS agreed upon a set of criteria for the selection, retirement, removal and suspension of measures applicable to all of its programs. The set contains core criteria and optional criteria. Program leads may apply the appropriate optional criteria to be used for specific programs.

Selection	Retirement	Removal	Suspension
Core Criteria	Core Criteria	Core Criteria	Core Criteria
1. Measure addresses an important condition/topic with a performance gap and has a strong scientific evidence base to demonstrate that the	1. Measure is owned by CMS and CMS will no longer maintain the measure	1. Measure is no longer used in a CMS program. If the measure is owned by CMS, CMS continues to maintain it even after removal.	1. Temporary in nature
	Optional Criteria	2. If another entity owns the measure, other	Optional Criteria
			2. Pending an ad hoc review, which could be

Selection	Retirement	Removal	Suspension
<p>measure when implemented can lead to the desired outcomes and/or more appropriate costs (i.e., NQF's Importance criteria).</p> <p>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).</p> <p>3. Promotes alignment with specific program attributes and across CMS and HHS programs</p> <p>4. Program measure <i>set</i> includes consideration for health care disparities</p> <p>5. Measure reporting is feasible.</p> <p>Optional Criteria</p> <p>6. Is responsive to specific program goals or statutory requirements.</p> <p>7. Enables measurement using measure type not already measured well (e.g., outcome, process, cost, etc.).</p> <p>8. Enables measurement across the person-centered episode of care,</p>	<p>2. No longer adds value commensurate with the cost of data collection and reporting</p> <p>3. Performance or improvement on a measure does not result in better outcomes</p> <p>4. Collection or public reporting of a measure leads to negative unintended consequences other than patient harm</p> <p>5. A measure does not align with current clinical guidelines or practice</p> <p>6. Measure performance is so high and unvarying that meaningful distinctions and improvements in performance can no longer be made</p> <p>7. The availability of a better measure, i.e.: <ul style="list-style-type: none"> — more broadly applicable (across settings, populations, or conditions) measure for the topic — more proximal in time to desired patient outcomes for the particular topic </p>	<p>payers/purchasers/programs using the measure are responsible for determining if the owner is continuing to maintain the measure.</p> <p>Optional Criteria</p> <p>3. Measure performance is so high and unvarying that meaningful distinctions and improvements in performance can no longer be made</p> <p>4. The availability of a better measure i.e.,: <ul style="list-style-type: none"> — more broadly applicable (across settings, populations, or conditions) measure for the topic; — more proximal in time to desired outcomes for the particular topic — more strongly associated with desired outcomes for the particular topic — more aligned with other CMS/HHS programs </p> <p>4. Collection or public reporting of a measure imposes undue burden</p> <p>5. The measure, as currently specified, cannot be reported</p>	<p>triggered by:</p> <ul style="list-style-type: none"> — significant evidence of changes in the evidence base of the measure or — significant evidence regarding unintended consequences potentially causing patient harm <p>3. The measure performance is high but there is concern that the measure performance might slip if it is retired</p>

Selection	Retirement	Removal	Suspension
<p>demonstrated by assessment of the person’s trajectory across providers and settings</p> <p>9. Program measure set promotes parsimony.</p>	<ul style="list-style-type: none"> — more strongly associated with desired patient outcomes for the particular topic — more aligned with other CMS programs 		

6. Measure Update

6.1 Introduction

The measure update process ensures that the Centers for Medicare & Medicaid Services (CMS) measures are updated as the code sets on which the measures rely are updated. Any comments and suggestions that were collected are also considered during measure update to determine if revision is needed beyond updating the codes.

The measure update process involves three parts:

- ◆ Gathering information that has been generated since the last review (the comprehensive reevaluation, measure update, or measure development—whichever occurred most recently).
- ◆ Recommending action.
- ◆ Approving and implementing that action.

6.2 Possible Outcomes

The possible outcomes in the measure update process:

- ◆ 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.

For measures not owned/maintained by a CMS measure contractor

The Measures Manager will include these measures in the measure inventory, along with anticipated dates when the steward will be reviewing the measures. As directed by CMS, the Measures Manager may monitor the Web site of the measure steward for updates. If there is no steward for a measure (contracted or non-contracted), CMS will decide whether resources can be allocated to conduct the measure update.

6.3 Deliverables

- ◆ An updated Measure Information form (MIF) showing all recommended changes to the measure.
- ◆ A document summarizing changes made, such as Release Notes if not included in the updated Measure Information Form
- ◆ NQF Annual Update online submission (whether or not any change was made to the measure)
- ◆ NQF submission documentation for any material changes¹ to the measure.

6.4 Procedure

Step 1: Complete Measures Management System orientation

As directed by CMS, the measure contractor may meet with the Measures Manager along with the Measures Manager's Contracting Officer Representative/Government Task Leader (COR/GTL) and the measure contractor's COR/GTL. The purpose of the orientation is to ensure that the measure contractor understands the Measures Management System requirements and measure reevaluation processes. The meeting will be a conference call and will be completed within two weeks after the award of the contract. This orientation is often waived for measure contractors who have experience conducting measure maintenance.

Step 2: Review the measure's code sets

Review the code sets used by the measure to determine:

- ◆ If new codes have been added to or deleted from the code set that may affect the measure.
- ◆ If codes have been changed so that their new meaning affects their usefulness within the measure.

If the measure has not been specified with ICD-10 codes, consider converting any ICD-9 codes to ICD-10. On January 16, 2009, the U.S. Department of Health and Human Services (HHS) released the 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.

Step 3: Review any unsolicited comments that have been submitted

If the feedback can be resolved with minimal change to the measure, consider doing so. If the feedback indicates a serious scientific concern with the clinical practice underlying the measure, incorporate an ad hoc review into the measure update. Refer to the Ad Hoc Review section for a detailed discussion of the procedure. If the feedback results in the need to change measure specifications, the measure contractor should evaluate the feasibility and impact of making the change during the review.

¹A 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.

Step 4: Review the results of the ongoing surveillance

The measure contractor is expected to have already been conducting ongoing surveillance of the measure's environment. This includes reviewing and managing comments on the measure, and reviewing literature pertinent to the measure. This surveillance may have resulted in an ad hoc review. Regardless, all new information should be considered during the measure update. Of particular concern is any evidence of unforeseen adverse consequences or any controversies that have arisen surrounding the measure.

Step 5: Conduct a limited review of the measure's performance

The limited review should include the following:

- ◆ National performance rate.
- ◆ State and regional performance rates.
- ◆ Variations in performance rates.
- ◆ Validity of the measure and its constituent data elements.
- ◆ Reliability of the measure and its constituent data elements.

Step 6: Determine the recommended disposition of the measure

In most cases, minimal research is necessary during the measure update process. The only anticipated input is:

- ◆ Updated code sets
- ◆ Unsolicited feedback from stakeholders
- ◆ Results of the ongoing surveillance
- ◆ Limited performance review

Based on the above information, revise the measures as needed, and review the measures against the measure evaluation criteria. Refer to Measure Evaluation During Maintenance Phase section for a discussion of the measure evaluation criteria.

Retention

If the measure continues to be suitable, document any specific coding changes or minor changes on the MIF. Describe a summary of the changes in the Release Notes/Summary of Changes section of the MIF or in a separate document. If an ad hoc review was conducted concurrently with the measure update, review the Measure Evaluation report and update it as needed. Minor changes include clarifying vague or inexact terminology or making graphic representations of the algorithm more accurate. 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. Any revision that changes the process of data collection, aggregation, or calculation is a material change.

Suspension

If the measure requires material revisions, assess whether those revisions are reasonable given the time and resources allocated by CMS for the measure's update. If the measure requires material revisions but those revisions are not feasible, document the recommendation that the measure be suspended pending an ad hoc or comprehensive reevaluation and the COR/GTL's approval to implement the revised measure. Include the specific barriers to retaining the measure in its current form. CMS may also choose to suspend the collection of a measure if the measure performance is high but there is concern that the measure performance might slip if it is retired.

Revision

If the measure requires material revision due to concerns based on scientific evidence, significant evidence regarding unintended consequences potentially causing patient harm, or another CMS concern, an ad hoc review should be incorporated into the measure update. Once the COR/GTL has agreed that an ad hoc review is needed, the measure contractor may need to consult the technical expert panel (TEP), gather public comments, or assist CMS in submitting the measure changes via appropriate rulemaking processes (e.g., Inpatient Prospective Payment System or Physician Fee Schedule). Refer to the Ad Hoc Review, Technical Expert Panel During Maintenance Phase, and Public Comment During Maintenance Phase sections for the processes.

Document the recommended changes in the Measure Information form (MIF). Incorporate those recommendations into the Measure Justification and Measure Evaluation report, along with any updates to the measure's strength or weakness resulting from the recommended changes.

Retirement

If an ad hoc review is incorporated into the measure update (refer to Revision above) and the measure is determined to be no longer suitable or cannot be made acceptable through reasonable revision (even with time allowed by suspending the measure), the measure should be retired. The criteria adopted by CMS to determine when a measure owned by CMS should be retired include when a measure:

- ◆ No longer adds value commensurate with the cost of data collection and reporting.
- ◆ Performance or improvement does not result in better outcomes.
- ◆ Collection or public reporting leads to negative unintended consequences other than patient harm.
- ◆ Does not align with current clinical guidelines or practice.
- ◆ Performance is so high and unvarying that meaningful distinctions and improvements in performance can no longer be made.
- ◆ Can be replaced by a better measure that is:
 - More broadly applicable (across settings, populations, or conditions) measure for the topic.
 - More proximal in time to desired patient outcomes for the particular topic.
 - More strongly associated with desired patient outcomes for the particular topic more aligned with other CMS programs.

Removal

A measure may be removed from 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. The criteria adopted by CMS to determine when a measure that is not owned by CMS should be removed include:

- ◆ Measure performance is so high and unvarying that meaningful distinctions and improvements in performance can no longer be made
- ◆ The availability of a better measure i.e.,:
 - more broadly applicable (across settings, populations, or conditions) measure for the topic;
 - more proximal in time to desired outcomes for the particular topic
 - more strongly associated with desired outcomes for the particular topic
 - more aligned with other CMS/HHS programs
- ◆ Collection or public reporting of a measure imposes undue burden
- ◆ The measure, as currently specified, cannot be reported

Step 7: Make a recommendation to the COR/GTL regarding the measure

Forward the recommendations to the COR/GTL, along with updated MIFs, Summary of Changes/Release Notes. If significant changes were made, an updated Measure Justification and Measure Evaluation report may be necessary.

Step 8: The COR/GTL reviews the recommendation for approval

The COR/GTL reviews the Measure Update documentation. If the recommendation is not approved, the COR/GTL documents the approved course of action and instructs the measure contractor as necessary. If the recommendation is approved, the COR/GTL notifies the measure contractor of the approved course of action.

Step 9: Implement the approved action

For measures that are proposed to be revised, suspended or retired, an evaluation of the impact of the decision on the program using the measure should be taken into consideration in developing the implementation plan. If implementation of the decision involves any regulatory or other schedules (e.g., rulemaking), include that schedule in the implementation planning.

Retain

Announce any minor changes to the measure (e.g., coding updates) through the usual communications modes for the project(s) in which the measure is used and arrange for any reprogramming that is necessary. Notify the Measures Manager of the changes to ensure that the CMS Measures Inventory is updated.

Revise

Announce the approved changes to the measure through the usual communications modes for the project(s) in which the measure is used and arrange for any necessary retraining or reprogramming. Calculate a new baseline, if appropriate. Notify the Measures Manager of the changes to ensure that the CMS Measures Inventory is updated.

Retire

Announce the effective date of the termination of data collection through the usual communications modes for the project(s) in which the measure is used. Notify the Measures Manager of the retirement to ensure that the CMS Measures Inventory is updated. Notify any other CMS contractors (e.g., Clinical Data Warehouse contractor) that may be impacted by the measure's retirement.

Suspend

Announce the terms of the suspension (e.g., data collection continues without public reporting, no further data collection) through the usual communications modes for the project(s) in which the measure is used and the effective date. Notify the Measures Manager of the suspension to ensure that the CMS Measures Inventory is updated. Notify any other CMS contractors (e.g., Clinical Data Warehouse contractor) that may be impacted by the suspension of data collection.

Remove

Announce the effective date of the termination of data collection through the usual communications modes for the project(s) in which the measure is used. Notify the Measures Manager of the retirement to ensure that the CMS Measures Inventory is updated. Notify any other CMS contractors (e.g., Clinical Data Warehouse contractor) that may be impacted by the measure's removal.

Step 10: Assist the COR/GTL in notifying NQF of the updated measure

After a measure is endorsed by NQF, CMS (as the measure steward) is required to submit a status report of the measure specifications to NQF on an annual basis. This report either affirms that the detailed measure specifications of the endorsed measure have not changed or, if changes have been made, the details and underlying reason(s) for the change(s). If changes occur to a measure at any time in the three-year endorsement period, the measure steward is responsible to inform NQF immediately of the timing and purpose of the changes. Some measure maintenance contracts may require updates to the measure more than once a year. In this case, the measure contractor may need to notify NQF of the changes each time they occur. If no changes are made, only one annual update is required.

NQF provides a standardized template for submission of annual measure maintenance that is prepopulated with measure information. Annual maintenance focuses on whether the measure remains current. CMS will direct NQF regarding the appropriate measure contractor to contact for the annual update. The measure contractor is responsible to prepare this report for NQF. The contractor must also obtain COR/GTL review and approval before submitting the report to NQF. NQF may conduct its own ad hoc review if the changes materially affect the measure's original concept or logic.

The contractor responsible for measure maintenance should be aware of when the annual update is due to NQF. The due date for the contractor's measures updates should be confirmed annually with NQF. The contractor should also inform NQF of any contact information changes, whenever these changes occur.

Step 11: Consider measures not owned/maintained by a CMS measure contractor

If the measure contractor responsible for implementing the measure for which CMS is not the measure owner (i.e., not ultimately responsible for maintaining the measure), then the measure contractor will be responsible for monitoring the measure owner's maintenance of the measure. This includes ensuring that the measure is revised periodically in response to updates in the underlying code sets (e.g., CPT, ICD-9, LOINC) and that the measure is reevaluated in a manner consistent with (though not necessarily identical to) the requirements of the Comprehensive Reevaluation section. The CMS measure contractor is also responsible for updating any CMS documentation of the measure to reflect changes made by the measure owner, and discussing those changes with CMS to ensure CMS wants to continue the use of the measure. Changes cannot be made to a measure that is copyright protected without the owner's consent. The measure contractor will also be responsible for ongoing surveillance of the literature addressing the measure and alerting the COR/GTL to possible issues.

7. Comprehensive Reevaluation

7.1 Introduction

The comprehensive reevaluation process ensures that the Centers for Medicare & Medicaid Services (CMS) measures continue to be of the highest caliber possible. By periodically reviewing the measures against the measure evaluation criteria, the measure contractor assists CMS in maintaining the best measures over time.

The comprehensive reevaluation process includes:

- ◆ Gathering information that has been generated since the last comprehensive reevaluation or since the measure’s development—whichever occurred most recently.
- ◆ Measure evaluation and recommending action.
- ◆ Approving and implementing that action.

This process assumes that the measure contractor has been conducting ongoing surveillance of the scientific literature and clinical environment related to the measure, including any clinical guidelines related to the measure.

7.2 Possible Outcomes

The outcomes possible when reviewing a measure:

- ◆ **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.

For measures not owned/maintained by a CMS measure contractor

The Measures Manager will include these measures in the measure inventory, along with anticipated dates when the steward will be reviewing the measures. As directed by CMS, the Measures Manager may monitor the Web site of the measure steward for updates. If there is no steward for a measure

(contracted or non-contracted), CMS will decide whether resources can be allocated to conduct the measure update.

7.3 Deliverables

- ◆ An updated Measure Information form (MIF) showing all recommended changes to the measure.
- ◆ A document summarizing changes made, such as Release Notes if not included in the updated Measure Information Form
- ◆ An updated Measure Justification documenting the environmental scan results, any new controversies about the measure, and any new data supporting the measure.
- ◆ An updated Measure Evaluation report reflecting the experience of the measure compared to the measure evaluation criteria.
- ◆ In the years when an endorsed measure is not being re-evaluated for continued endorsement, measure stewards will submit a status report of the measure specifications to NQF
- ◆ NQF endorsement maintenance documentation (at the time of NQF's next scheduled three-year maintenance review).

7.4 Procedure

Step 1: Complete Measures Management System Orientation

As directed by CMS, the measure contractor may meet with the Measures Manager along with the CMS staff members overseeing the Measures Manager and the measure contractor. The purpose of the orientation is to ensure that the measure contractor understands the Measures Management System requirements and measure reevaluation processes. The meeting can be face-to-face or via conference call and must be completed within two weeks after the award of the contract. At the discretion of the Contracting Officer Representative/Government Task Leader (COR/GTL), this orientation may be waived for measure contractors who have experience conducting measure maintenance.

Step 2: Develop a work plan

As directed by the measure contractor's scope of work, a work plan for the measure(s) 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 required measurement system processes and has a strategy for executing them. The measure contractor will refer to the contract scope of work for the date when the work plan is due.

When developing the work plan, two other schedules should be considered:

- ◆ The rulemaking cycle for any regulatory process governing the measure set in question.
- ◆ The National Quality Forum (NQF) measure maintenance schedule.

Step 3: Obtain current performance data on each measure

Performance information includes, but is not limited to:

- ◆ Current aggregate national and regional measurement results.
- ◆ Measurement results trended across the years since the measure's initial implementation.
- ◆ The current distribution of measurement results by provider types (e.g., rural vs. urban, for-profit vs. nonprofit, facility bed size, etc.).
- ◆ Analysis of the measure's reliability, stability, and validity since implementation.
- ◆ The results of audits and data validation activities.
- ◆ Analysis of any disparities in quality of care based on race, ethnicity, age, socioeconomic status, income, region, gender, primary language, disability, or other classifications.
- ◆ Analysis of unintended consequences that have arisen from the use of the measure.
- ◆ Validation and analysis of the exclusions, including, but not limited to:
 - Analysis of variability of use.
 - Implications of rates.
- ◆ Other performance information that CMS has collected or calculated, as available.

Step 4: Analyze the ongoing feedback received

Review the feedback that has been received since the measure was last evaluated (either the initial evaluation or the last comprehensive reevaluation, whichever is most recent). Identify any changes or concerns that are necessary to respond to the feedback.

Step 5: Summarize the findings of the ongoing environmental scan

The ongoing environmental scan should focus on information published or otherwise available since the last time the measure was evaluated.

Refer to the Information Gathering During Maintenance Phase section for the standardized process. At a minimum this synthesis should include:

- ◆ Changes to clinical guidelines on which the measure is based.
- ◆ Relevant studies that might change clinical practice, which in turn might affect the underlying assumptions of the measure.
- ◆ Relevant studies that document unintended consequences of the measure.
- ◆ Relevant studies that document continued variation or gaps in the care being measured.
- ◆ Technological changes that might affect how data is collected, calculated, or disseminated.
- ◆ Similar measures based on their structure, clinical practices, or conditions that could offer an opportunity for harmonization or might serve as replacement measures.
- ◆ Relevant information gathered from the Technical Expert Panel (TEP) or interviews with subject matter or measurement experts.
- ◆ Reevaluation of the business case supporting the measure.

Step 6: Convene a TEP

Refer to the Technical Expert Panel During Maintenance Phase section for the standardized process. Involving TEP members who have reviewed the measure(s) previously is recommended. The TEP can be convened by either teleconference or a face-to-face meeting. Present the results of the environmental scan, literature review, empirical data analysis of the measure performance data, and analysis of ongoing feedback received. Develop recommendations on the disposition of the measure using the measure evaluation criteria. Refer to the Measure Evaluation During Maintenance Phase section for a discussion of the measure evaluation criteria.

Summarize the TEP's recommendations on the MIF and in the TEP report.

Step 7: Synthesize the results of the preceding steps

For each measure, synthesize and compile the information gathered in the steps above using the measure evaluation criteria and the Measure Evaluation report to identify the strengths and weaknesses of each measure.

Document the findings on the Measure Information form (MIF), the Measure Justification form, or the Measure Evaluation report as appropriate, including any changes to the measure's ratings relative to the evaluation criteria.

If the measure has not been specified with ICD-10 codes, consider converting any ICD-9 codes to ICD-10. On January 16, 2009, the U.S. Department of Health and Human Services (HHS) released the 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.

Step 8: Identify and document changes that will be recommended

All changes to measure specifications should be documented on the MIF in the Release Notes/Summary of Changes section or in a separate document. Any material changes should be identified and the purpose of the changes explained. 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.

Step 9: Determine the preliminary recommended disposition of the measure

Refer to the Measure Evaluation During Maintenance Phase section for a discussion of the measure evaluation criteria, which form the basis for the disposition decision for each measure.

Retirement

If CMS's measure performance data indicates that most providers are performing at a high level, or if conditions (i.e., the characteristics of the measure rated by the measure evaluation criteria) have changed sufficiently to render the measure no longer desirable, consider retiring the measure.

Retirement refers to the termination of data collection and reporting of the measure for the foreseeable future. The criteria adopted by CMS to determine when a measure owned by CMS should be retired include when a measure:

- ◆ No longer adds value commensurate with the cost of data collection and reporting.
- ◆ Performance or improvement does not result in better outcomes.
- ◆ Collection or public reporting leads to negative unintended consequences other than patient harm.
- ◆ Does not align with current clinical guidelines or practice.
- ◆ Performance is so high and unvarying that meaningful distinctions and improvements in performance can no longer be made.
- ◆ Can be replaced by a better measure that is more:
 - Broadly applicable (across settings, populations, or conditions) measure for the topic.
 - Proximal in time to desired patient outcomes for the particular topic.
 - Strongly associated with desired patient outcomes for the particular topic more aligned with other CMS programs.

For the benefit of the measure set, retiring one measure may offer an opportunity to add another measure. The decision to retire a particular measure should, however, be assessed based on the merits of the measure itself before considering if another measure is better.

Though these criteria were published to apply to specific programs, they should be included when considering the retirement of any measure.

Suspension

CMS may choose to suspend the collection of a measure if the measure performance is high but there is concern that the measure performance might slip if it is retired.

Revision

Having decided that the measure is not ready for retirement and is still the best measure available, determine if the current limitations to the measure can be improved according to the measure evaluation criteria. Examples of such improvements may include:

- ◆ Changing the risk adjustment methodology.
- ◆ Including additional types of surgery in the denominator of a surgery measure.

If the measure cannot be revised sufficiently to make it acceptable according to the measure evaluation criteria, recommend that the measure be retired.

NQF requires that any material changes made to NQF-endorsed measures be reported immediately. Refer to the National Quality Forum Endorsement During Maintenance Phase section and the NQF Web site (<http://www.qualityforum.org>) for more information.

The measure revision(s) should be documented on the Measure Information Form, located in the Measure Development section of Volume 1, detailed in the Technical Specifications section of the MIF

and summarized in the Release Notes/Summary of Changes. Supporting material is documented in the Measure Evaluation report (refer to Measure Evaluation During Maintenance Phase section for more information).

Retention

If the measure is not to be retired or materially revised, it is retained. (Any revision that changes the process of data collection, aggregation, or calculation is a material change.) Recommend to the COR/GTL that the measure be retained, citing any coding updates that are necessary.

Removal

A measure may be removed from 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. The criteria adopted by CMS to determine when a measure that is not owned by CMS should be removed include:

- ◆ Measure performance is so high and unvarying that meaningful distinctions and improvements in performance can no longer be made.
- ◆ A better measure is available, such as one that is more:
 - Broadly applicable (across settings, populations, or conditions) measure for the topic.
 - Proximal in time to desired outcomes for the particular topic.
 - Strongly associated with desired outcomes for the particular topic.
 - Aligned with other CMS/HHS programs.
- ◆ Collection or public reporting of a measure imposes undue burden.
- ◆ The measure, as currently specified, cannot be reported.

Step 10: Obtain preliminary approval from the COR/GTL to release the recommended measure disposition for public comment

The COR/GTL provides approval to release the measure for public comment.

If the measure contractor's recommendation is not approved for such release, the COTR/GTL documents the action to be taken and notifies the measure contractor of the approved alternative. Refer to the Public Comment During Maintenance Phase section for the process.

Step 11: Obtain public comment on the measure

If the comprehensive reevaluation results in a recommendation to retain the measure with only minor changes, skip to the next step.

If the measures in question fall within the rulemaking processes (e.g., Inpatient Prospective Payment System or Physician Fee Schedule), the public comment solicited through the rulemaking process satisfies this step.

The results of the comprehensive reevaluation, including the preliminary recommended disposition and any revised or replacement measures will be posted on the dedicated CMS Web site for public comment.

Refer to the Public Comment During Maintenance Phase section for the standardized process.

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 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>. 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 12: Refine the recommended measure reevaluation dispositions

Analyze the comments received and refine the recommended disposition of the measure. Document the recommendation and rationale for it in the MIF and measure Evaluation report, as appropriate, including any relevant public comments, and send the completed form to the COR/GTL for approval.

Any further revisions of the technical specifications resulting from the comments received should be documented on an updated MIF and summarized in the Release Notes/Summary of Changes section of the MIF.

Step 13: Obtain the COR/GTL's approval

The COR/GTL provides approval for the recommendation and notifies the measure contractor. If the measure contractor's recommendation is not approved, the COR/GTL documents the action to be taken and notifies the measure contractor of the alternative action that has been approved.

Step 14: Implement the approved action

For measures that are proposed to be revised, suspended or retired, an evaluation of the impact of the decision on the program using the measure should be taken into consideration in developing the implementation plan. If implementation of the decision involves any regulatory or other schedules (e.g., rulemaking), include that schedule in the implementation planning.

Retain

Announce any minor changes to the measure (e.g., coding updates) through the usual communications modes for the project(s) in which the measure is used and arrange for any reprogramming that is necessary. Notify the Measures Manager to ensure that the CMS Measures Inventory is updated.

Revise

Announce the approved changes to the measure through the usual communications modes for the project(s) in which the measure is used and arrange for any necessary retraining or reprogramming. Calculate a new baseline, if appropriate. Notify the Measures Manager to ensure that the CMS Measures Inventory is updated.

Retire

Announce the effective date of the termination of data collection through the usual communications modes for the project(s) in which the measure is used. Notify the Measures Manager to ensure that the CMS Measures Inventory is updated. Notify any other CMS contractors (e.g., Clinical Data Warehouse contractor) that may be impacted by the measure's retirement.

Suspend

Announce the terms of the suspension (e.g., data collection continues without public reporting, no further data collection) through the usual communications modes for the project(s) in which the measure is used and the effective date. Notify the Measures Manager of the suspension to ensure that the CMS Measures Inventory is updated. Notify any other CMS contractors (e.g., Clinical Data Warehouse contractor) that may be impacted by the suspension of data collection.

Remove

Announce the effective date of the termination of data collection through the usual communications modes for the project(s) in which the measure is used. Notify the Measures Manager to ensure that the CMS Measures Inventory is updated. Notify any other CMS contractors (e.g., Clinical Data Warehouse contractor) that may be impacted by the measure's removal.

Step 15: Assist in notifying NQF of the updated measure

The COR/GTL will notify NQF of all relevant activities and changes to the measure or instruct the measure contractor to do so on behalf of CMS. The measure contractor is expected to assist the COR/GTL in preparing any documents necessary for NQF and in responding to any questions or comments from NQF.

Step 16: Assist in the NQF endorsement maintenance review

Every three years, endorsed measures are re-evaluated against the NQF's Measure Evaluation Criteria and are reviewed alongside newly submitted (but not yet endorsed) measures. This head-to-head comparison of new and previously endorsed measures fosters harmonization and helps ensure NQF is endorsing the best available measures.

NQF will formally notify the CMS of the upcoming expiration of a measure's NQF endorsement six months prior to that date. Then the Measure Manager or the COR/GTL will notify the appropriate measure contractor. The measure contractor should be aware of the NQF maintenance cycle, and if they have not received notice of a maintenance review, the measure contractor should contact NQF or their COR/GTL.

NQF will provide a standardized online submission template for the three-year maintenance review. The form will be prepopulated with information from the most recent submission. CMS will direct NQF regarding the appropriate measure contractor contact. The measure contractor is responsible for updating the information in the form and for obtaining COR/GTL review and approval before submission to NQF.

The three-year maintenance review form documents the review of the current evidence and guidelines and provides information about how the measure still meets the criteria for NQF endorsement. The measure contractor will use information from the most recent Comprehensive Reevaluation, subsequent Measure Updates and ongoing surveillance to complete the NQF three-year maintenance Form. The measure contractor obtains COR/GTL approval and submits to NQF by the required date.

8. Ad Hoc Review

8.1 Introduction

The ad hoc review process ensures that the Centers for Medicare & Medicaid Services (CMS) measures remain balanced between the need for measure stability and the reality that the measure environment is constantly shifting. However ad hoc review should be reserved to only those instances where new evidence indicates significant revision may be required of the greatest significance.

Ad hoc review specifically does not include the process of adapting or harmonizing a measure for use with a broader or different population.

Structure of the ad hoc review

The ad hoc review process includes three primary subparts:

- ◆ Determining if an ad hoc review should be conducted.
- ◆ Conducting the review and recommending an outcome.
- ◆ Approving and implementing the approved outcome.

Trigger for an ad hoc review

The potential ad hoc review begins when the measure contractor becomes aware of evidence that may have a significant, adverse effect on the measure or its implementation. The evidence may become known through the measure contractor's ongoing surveillance of the scientific literature relevant to the measure, from the Measures Manager, CMS, stakeholders, or other sources.

If the measure is National Quality Foundation (NQF) endorsed, NQF may have received a request for an ad hoc review and may have contacted the steward CMS. CMS may request the measure contractor to investigate the situation and conduct its own ad hoc review even if NQF has declined to conduct an ad hoc endorsement review. If NQF has decided to conduct an ad hoc endorsement review, the measure contractor will be asked to assist CMS in assessing the situation and to provide information for NQF review. NQF ad hoc reviews may be initiated at the request of CMS for specific situations, such as the need to significantly change measure specification outside of the normal maintenance cycle. For example, NQF has required CMS to harmonize a measure before the next maintenance review; however, CMS needs the revised measure prior to that time due to program or legislative requirements, and must use an endorsed measure.

CMS reserves the right to conduct an ad hoc review for any reason, at any time, on any measure. Nothing in this Blueprint is intended to limit the options CMS may exercise.

Deferring an ad hoc review

Postpone an ad hoc review to the next scheduled review if that is reasonable. The timing of the ad hoc review will be influenced by the presence of any accompanying patient safety concerns associated with the changes to the endorsed measure. If the measure will be updated or reevaluated in the near

future, the information received should be incorporated into that update or reevaluation. For example, if the measure is due for a comprehensive reevaluation or a measure update within the next 120 days, the information should be referred to the team conducting the review and that team should incorporate the ad hoc review process into its work.

Because measures are used in particular programs which may have their own schedules (e.g., hospital measures are governed by the inpatient prospective payment system rulemaking schedule requirements), a decision may take some time to be implemented in all the programs using a given measure.

8.2 Possible Outcomes

The following outcomes are possible after conducting an ad hoc 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 publicly report a measure. Data collection and submission may continue, as directed by CMS. (This option may be used as an interim decision by CMS when awaiting final resolution of an issue.)
- ◆ **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.

8.3 Deliverables

- ◆ Updated Measure Information form (MIF), if the ad hoc review results in changes to the measure specifications.
- ◆ Updated Measure Justification, reflecting the new information that triggered the review, any additional information used in the decision-making process, and the rationale for the outcome of the review.
- ◆ Updated Measure Evaluation report, if the review resulted in a change to the measure’s strengths and/or weaknesses.

8.4 Procedure

For measures not owned/maintained by a CMS measure contractor

If the CMS contractor responsible for implementing a measure is not the measure owner (that is, not the steward or ultimately responsible for maintaining the measure), the CMS contractor remains responsible to monitor the maintenance done by the owner. This includes ensuring that the measure is updated periodically in response to changes in the underlying code sets (e.g., CPT, ICD-9-CM, LOINC) and is reevaluated in a manner consistent with the Blueprint. The CMS measure contractor will also be responsible for ongoing surveillance of the literature addressing the measure and alerting the Contracting Officer Representative/Government Task Leader (COR/GTL) to possible issues. In the absence of such a CMS measure contractor and at the request of CMS, the Measures Manager may perform these functions.

If a significant concern is identified with a measure for which CMS is not the steward, the measure contractor responsible for monitoring the measure should bring the matter to the attention of the COR/GTL to determine what action, if any, is necessary. CMS may contact the steward to determine if the steward is aware of the concern and what action is being taken. If the measure is NQF-endorsed, CMS may consider requesting NQF to conduct an ad hoc maintenance review. CMS has the option of suspending data collection pending the outcome of any action by the steward and NQF, or CMS may choose to remove the measure from the program.

For measures owned/maintained by a CMS measure contractor, follow the steps below

Step 1: Determine if the concern is significant

Only the strongest concerns will result in an ad hoc review. The determination of “significant” issues should be considered initially by the measure contractor monitoring the measure and may involve the technical expert panel (TEP) that was involved with the measure most recently. If the contractor does not have access to the TEP, then the contractor may contact a professional association closely associated with the measure for input regarding the significance of the issue raised.

If the experts determine that the issue is not significant, the issue should be documented for consideration at the next scheduled review.

If the experts determine that the issue is significant, or if they cannot agree on its significance, the contractor should notify CMS of the situation and propose conducting a full ad hoc review (the remaining steps below). If the measure maintenance contractor is different from the contractor monitoring the measure, the measure maintenance contractor should be responsible for the review.

Step 2: Conduct an environmental scan

Unlike environmental scans conducted during measure development, ongoing surveillance, or comprehensive reevaluation, the scan done for an ad hoc review is limited to information directly

related to the issue resulting in the review. Not all aspects of the measure must be investigated, only the aspect that generated concern.

Conduct a literature review to determine the extent of the issues involved 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 in the search. Whenever possible, include the electronic versions of articles or publications when submitting recommendations to the COR/GTL.

- ◆ Specifically, attention should be paid to evidence supporting or contradicting the issue.
- ◆ Include 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.
- ◆ Search for other sources of information—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.
- ◆ Use systematic literature reviews¹ to assess the overall strength of the body of evidence if available for the issue of interest.
- ◆ Resources:
 - Institute of Medicine report: Finding What Works in Health Care Standards for Systematic Reviews², published March 23, 2011.
 - NQF Report: Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement and Importance to Measure and Report³, published January 2011.

Step 3: Consult with the experts, especially the TEP

The technical expert panel that assisted in the most recent comprehensive reevaluation or measure development may be consulted, if available.

If the issue generating the concern relates to clinical guidelines, the organization responsible for the guidelines may be consulted regarding its plans for updating the guidelines or issuing interim guidelines.

The professional organization most closely related to the measure may also be consulted.

The experts (TEP, guideline writers, or professional organization) should be queried about:

- ◆ The significance of the issue raised, in order to confirm the initial determination.

¹A 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.

²<http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx>

³http://www.qualityforum.org/Measuring_Performance/Improving_NQF_Process/Evidence_Task_Force.aspx

- ◆ The risk of patient harm resulting from continued use of the measure, including harm caused by unintended consequences.
- ◆ The feasibility of implementing a revision to the measure, including both costs and time necessary to do so.

Step 4: Determine if the concern involves patient safety

If the clinical practice underlying the measure is causing harm to the patients, the measure should be revised, suspended, or retired. This includes harm caused by unintended consequences of the measure. If the measure revision is not feasible (refer to the next step), the measure should be suspended or retired.

If no such harm is envisioned, the measure may be retained as it is until the next scheduled review unless CMS requires the review before that time.

Step 5: Determine if it is feasible to change the measure

The feasibility of changing a measure should include consideration of the cost of resources associated with data collection, measure calculation, and reporting systems, including those requiring updates to vendor systems. Depending on the resources available and the time involved in making the changes necessary, the measure may be either revised immediately or suspended until the systems can be updated with the measure's updated specifications.

Step 6: Recommend a course of action to the COR/GTL

Depending on the findings of Steps 1-3, the recommendation may be:

- ◆ Retain—If the issue is not significant or if there is no patient harm resulting from the measure.
- ◆ Revise—If the issue is significant, there is patient harm likely from continued use of the measure, and the revision can be done quickly.
- ◆ Suspend—If the issue is significant, and revision cannot be completed quickly.
- ◆ Retire—If the issue is significant, there is patient harm likely from continued use of the measure, and revision is not feasible.
- ◆ Remove—If the measure is no longer appropriate for a CMS program. 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.

Step 7: Obtain the COR/GTL's approval for the final recommendation

Submit the recommendation along with supporting documentation and the updated MIF and MER (if recommending immediate revision or suspension until revision is possible) to the COR/GTL. If the COR/GTL does not approve the proposed action, the COR/GTL will advise the measure contractor of the approved course.

Step 8: Implement the approved action

- ◆ Retain—Document that the measure is being retained. Notify the Measures Manager of the outcome of the ad hoc review.
- ◆ Revise—Announce the approved changes to the measure and arrange for any necessary retraining or reprogramming. Notify the Measures Manager of the outcome of the ad hoc review.
- ◆ Retire or Remove—Announce the effective date of the termination of data collection. Notify the Measures Manager of the outcome of the ad hoc review. Notify any other CMS contractors (e.g., Clinical Data Warehouse contractor) that may be impacted by the measure’s retirement.
- ◆ Suspend—Announce the terms of the suspension (no further reporting, no further data collection, etc.) and the effective date. Notify the Measures Manager of the outcome of the ad hoc review. Notify any other CMS contractors (e.g., Clinical Data Warehouse contractor) that may be impacted by the suspension of data collection.

Step 9: Assist the COR/GTL in notifying NQF of the updated measure

Refer to the NQF Web site for the current NQF measures maintenance policies.

The COR/GTL will notify NQF of all relevant activities and changes to the measure. The measure contractor will be available to answer NQF questions about the revision and/or the ad hoc review process.

As noted above, NQF also has an ad hoc review process. An ad hoc review may be conducted on an endorsed measure at any time if the evidence supporting the measure has changed, implementation of the measure results in unintended consequences, or material changes have been made to the measure. Ad hoc reviews can be requested at any time by any party; NQF reviews the request and initiates a review as long as there is adequate evidence to justify the review. When requesting an ad hoc review, requestors are asked indicate under which criterion they are requesting the ad hoc review and to provide in writing adequate evidence to justify the review.

The ad hoc review process follows a shortened version of the Consensus Development Process and includes a call for nominations for technical experts, review by the expert panel, a public and Member comment period for no less than 10 days, review by the CSAC, ratification by the NQF Board of Directors, and an appeals period.

Measure contractors will be responsible to assist CMS in responding to NQF requests for ad hoc review as a part of maintenance support. Contractors should notify NQF and provide updated measure specifications once CMS has approved them. Should CMS decide to remove a measure after ad hoc review, CMS will notify NQF. Measures that remain endorsed after an ad hoc review are still subject to the original maintenance cycle.

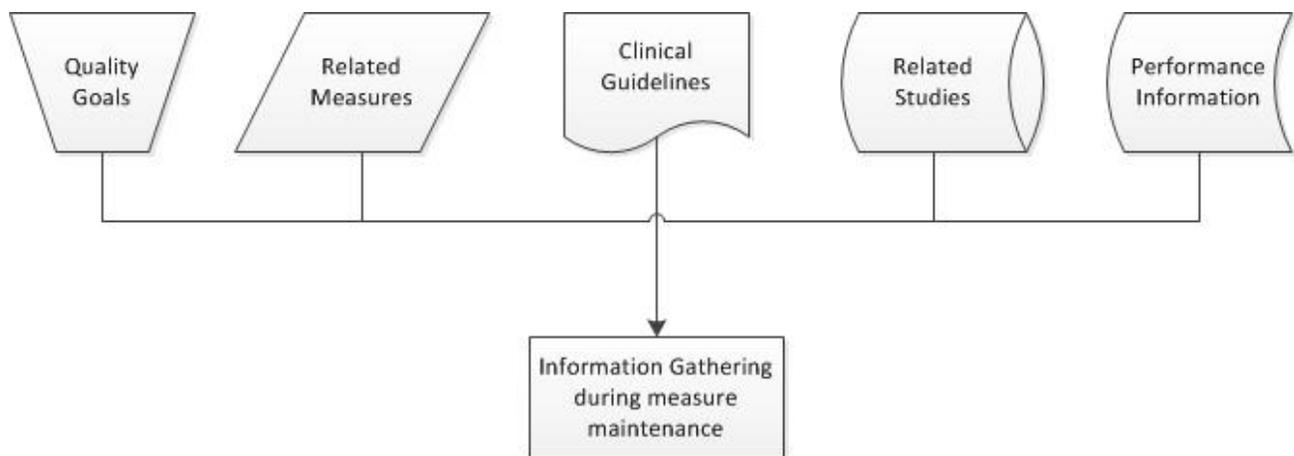
NQF currently does not have a form for the ad hoc review. NQF will inform the measure steward of the request and evidence presented by the requestor and will request the steward to respond.

9. Information Gathering During Maintenance Phase

9.1 Introduction

The purpose of this section is to provide guidance on gathering, organizing and documenting information during measure maintenance. Information gathering is a broad term that includes an environmental scan (literature review, clinical performance guidelines search, and search for related measures), measure performance information and empirical data analysis. These activities are conducted to obtain information that will guide the review and update of the measure justification, update of the business case, and harmonization with related measures, as well as documentation of the measure’s continued usefulness as a tool for quality improvement. If the measure is endorsed by NQF, this information will be used to support the measure during endorsement maintenance reviews and to complete the NQF measure maintenance form. This section describes the various sources of information that can be gathered as well as instructions for documenting and analyzing the collected information. Figure 9-1 below illustrates the items that should be researched during the information gathering process for measure maintenance.

Figure 9-1 Information Gathering



Just as information gathering is the first step in the development of new measures, it is an important step in measure maintenance. Good information gathering will ensure maintenance of a significant knowledge base that includes the measure’s recent performance, quality goals, the strength of scientific evidence (or lack thereof) pertinent to the topics/conditions of interest, and information for ensuring a continued business case for the measure. It will also update evidence of general agreement on the quality issues pertinent to the topics/conditions of interest and identify diverse or conflicting views. As in measure development, the four measure evaluation criteria—importance, scientific acceptability of measure properties, feasibility, and usability—will serve as a guide for conducting information gathering activities. 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 useful when submitting the measure for National Quality Forum (NQF) endorsement.

9.2 Deliverables

The deliverables necessary to document work conducted during the information-gathering phase of measure maintenance.

- ◆ Updated Measure Information form
- ◆ Measure Justification form

9.3 Procedure

Follow the steps below and document your progress using the Measure Justification form. The steps are not meant to be sequential but can be undertaken simultaneously.

Step 1: Conduct an environmental scan

Gather information found during the ongoing surveillance that is part of monitoring the use of the measure. A more comprehensive environmental scan should be conducted during the comprehensive review to ensure that all pertinent information has been found. This includes literature review, clinical practice guidelines review, and search for existing and related measures.

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 reviews¹ 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:
 - 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

Clinical practice guidelines review

Search for any new or updated clinical practice guidelines applicable to the measure topic written since the last comprehensive reevaluation if one was done, or since development. A good resource is the National Guideline Clearinghouse, located online at <http://www.guideline.gov/>. Clinical practice guidelines vary in how they are developed. Guidelines developed by American national physician organizations or federal agencies are preferred. Preferred guidelines are those developed with balanced representation beyond one specialty group and with full disclosure of biases and how they are addressed. The evidence underlying the guideline recommendation must be accessible. 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 and 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 evaluating clinical guidelines:

- ◆ National Guidelines Clearinghouse located at <http://www.guideline.gov/>.

¹A 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 *Clinical Practice Guidelines We Can Trust*. Released in March 2011, this report is available online at <http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust.aspx>.
- ◆ Physician Consortium for Performance Improvement® (PCPI) Position Statement *The Evidence Base Required for Measures Development* is a good resource for evaluating the guidelines and their acceptability in measure development. This document was approved in June 2009 and is available online at <http://www.ama-assn.org/resources/doc/cqi/pcpi-evidence-based-statement.pdf>.

Search for existing and related measures

Two objectives are:

- ◆ To identify measures with which the measure under review can and/or should be harmonized.
- ◆ To identify measures that may assess the same topic (and possibly, but not necessarily the same leverage point) in a way that contributes more to the overall measure set than the measure under review.

Keep the search parameters broad to obtain an overall understanding of a variety of measures that can be candidates for harmonization activities. 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. Table 9-1 outlines other measure search parameters that can aid in harmonization efforts.

Table 9-1 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

Use a variety of databases and sources to search for existing and related measures. Below are links to available sources²:

- ◆ 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/Home.aspx>

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

- ◆ American Medical Association-Physician Consortium for Performance Improvement—
<http://www.ama-assn.org/apps/listserv/x-check/qmeasure.cgi?submit=PCPI>

Step 2: Obtain and evaluate current performance data on measure under review.

Since the measure has already been available for implementation and data collection, the measure contractor should gather the performance information of the measure and analyze that to help with the comprehensive reevaluation of the measure. Analysis of the performance data should include but not be limited to:

- ◆ Current aggregate national and regional measurement results.
- ◆ The measurement results trended across the years since the measure’s initial implementation.
- ◆ The current distribution of measurement results by provider types (e.g., rural vs. urban, for-profit vs. nonprofit, facility bed size, etc.).
- ◆ Analysis of the measure’s reliability, stability, and validity since implementation.
- ◆ The results of audits and data validation activities.
- ◆ Analysis of any disparities in quality of care based on race, ethnicity, age, socioeconomic status, income, region, gender, primary language, disability, or other classifications.
- ◆ Analysis of unintended consequences that have arisen from the use of the measure.
- ◆ Validation and analysis of the exclusions, including, but not limited to:
 - ◆ Analysis of variability of use.
 - ◆ Implications of rates.
- ◆ Other performance information that CMS has collected or calculated, as available

Step 3: Evaluate existing or related measures for possible harmonization

If there are related measures, consider harmonization issues related to similar data elements (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.). Consult with the Measures Manager to review specifications to identify opportunities for further harmonization. Refer to the Harmonization During Maintenance Phase section for issues related to harmonization considerations. Table 9-2 provides guidance on when and where harmonization may be appropriate.

Table 9-2 Harmonization Decisions

	Harmonization Issue	Action
Numerator: Same measure focus Denominator: Same target population	Competing measures	<ul style="list-style-type: none"> ◆ Retain existing measure, or ◆ Replace existing measure with the similar measure identified; justify the need for change.

	Harmonization Issue	Action
Numerator: Same measure focus Denominator: Different target population	Related measures	<ul style="list-style-type: none"> ◆ Harmonize on measure focus, or ◆ Justify differences.
Numerator: Different measure focus Denominator: Same target population	Related measures	<ul style="list-style-type: none"> ◆ Harmonize on target population, or ◆ Justify differences.
Numerator: Different measure focus Denominator: Different target population	Unique measures	<ul style="list-style-type: none"> ◆ No action necessary, or ◆ Recommend the development of a new measure.

Harmonization should not result in inferior measures—measures should be based on the best measure concepts and ways to measure those concepts. There should be no presupposition that either the measure being reevaluated or the measure that was identified as “similar” is superior. The components of measure or measures that were identified should be evaluated in comparison to the measure being evaluated. Harmonization should eliminate unintended differences among related measures. There are circumstances where differences in measures should be expected: conceptually (e.g., difference in evidence for different patient populations) or technically (e.g., implementation in different settings with different data). In these cases, the measure contractor should explain the justification for not harmonizing with other existing measures. When there is a decision not to harmonize measures, the value of the different conceptualizations and technical specifications must outweigh the burden imposed. For further details on harmonization, please refer to the Harmonization During Maintenance Phase section.

Competing measures—Measures found during the search for measures that are intended to address both the same focus and the same target population can be considered “competing” measures. Use the measure evaluation criteria to determine if the measure undergoing reevaluation or the “competing” measure that was identified should be recommended for use. Listed below are considerations for evaluating the measures.³

Importance

Competing measures generally will be the same in terms of impact, opportunity for improvement, and evidence for the focus of measurement.

Scientific acceptability of measure properties

- ◆ Untested measures cannot be considered superior to tested measures.
- ◆ Compare and identify differences in specifications and methods.

³This list of considerations was adapted from *National Quality Forum. Memorandum to the Consensus Standards Approval Committee (CSA)*, from Helen Burstin and Karen Pace; Subject: *Usability and Feasibility*. December 2, 2010.

- ◆ Measures with the broadest application (settings, target population) are preferred.
- ◆ Compare ratings on reliability and validity and the overall criterion rating.

Usability

Compare ratings, all else being equal:

- ◆ Measures that are in use and publicly reported are preferred.
- ◆ Measures with the broadest use (e.g., settings, numbers of entities reporting performance results) are preferred.

Feasibility

Compare ratings all else being equal:

- ◆ Measures based on data from electronic sources are preferred.
- ◆ Measures that are freely available are preferred.

Step 4: Summarize the evidence in the Measure Justification form

Analyze the literature review results and the guidelines found and organize the evidence to support as many of the five measure evaluation criteria as possible. Update the information in the Measure Justification form for measures being recommended for continuation (either retention or revision). For measures not being recommended for continuation (i.e., recommended for retirement), document the reasons.

Step 5: Update the business case

Update the business case for measure under review and document it in the “Importance” section of Measure Justification form. Most of the information needed to update the business case will have been obtained through information gathering though the measure contractor may need to search for additional information.

There may be relevant topics which are beneficial for society in general or of ethical value but may not be associated with financial savings to CMS. The benefits derived from interventions promoted by these 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.

- ◆ Incidence/prevalence of condition in the population included in the measure.
- ◆ 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.

Step 6: 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 can be part of a report of the measure maintenance activities. It may include, but not be limited to:

- ◆ The methods used to gather information.
- ◆ The results of each activity conducted.
- ◆ A summary of the information gathered using the Information Gathering report (found in the Information Gathering section of Volume 1).

10. Technical Specifications During Maintenance Phase¹

10.1 Introduction

The purpose of this section is to provide guidance to the measure contractor to ensure that measures developed for the Centers for Medicare & Medicaid Services (CMS) have complete technical specifications that are detailed and precise. The final technical specifications provide the comprehensive details that allow the measure to be collected and implemented consistently, reliably, and effectively.

The process of updating measure specifications is iterative and occurs throughout the measure maintenance processes. Refer to the Measure Maintenance section for a broader discussion of the types of maintenance.

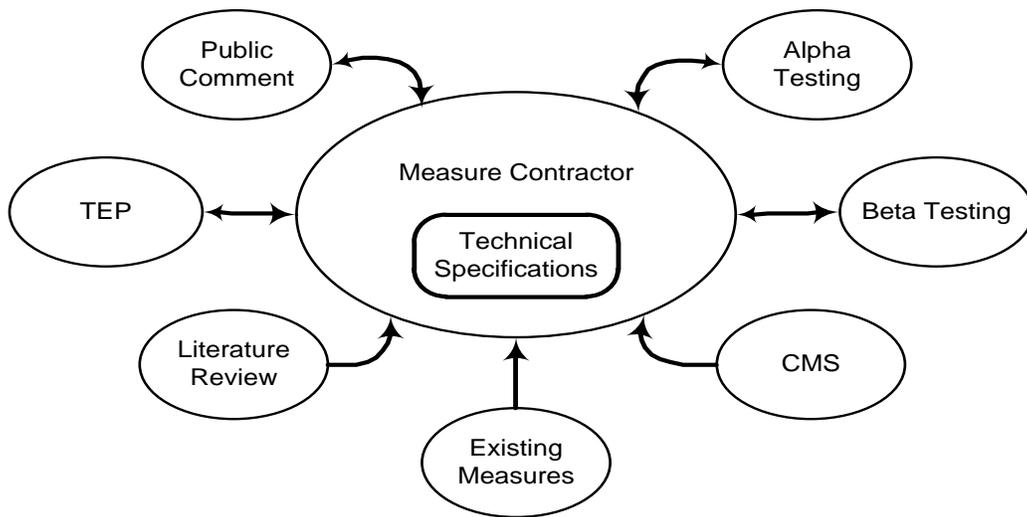
In the ongoing environmental scan (refer to the Measure Production and Monitoring section) and the information gathering stage of measure maintenance, the measure contractor identifies issues with the measure that may need to be addressed during the comprehensive reevaluation or an ad hoc review. The contractor may also find other measures with which the CMS measure should be harmonized. The measure contractor should work with its technical expert panel (TEP) to determine the recommended disposition of the measure. Depending on the findings of the information gathering, the TEP will recommend retaining, retiring, revising or suspending the measure. The measure contractor will review those recommendations and make its own recommendation regarding the disposition 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)—or equivalent document containing the same elements, if approved by the COR/GTL—is updated as the measures are updated, while the Measure Justification form will be updated with the findings from information gathering. If substantial changes are being considered, these changes may require testing of a limited scope. Refer to the Measure Testing During Maintenance Phase section for an in-depth discussion of how to conduct testing.

Figure 10-1 illustrates the inputs to technical specifications.

¹The direction provided in this section is partially based on guidance from the National Quality Forum and in some instances the verbiage remains unchanged to preserve the intent of the original documents.

Figure 10-1 Factors influencing the updating of the measure technical specifications



As required during measure development, the measures must be precisely specified and with sufficient details to be distinguishable from other measures and to enable consistent implementation across providers.

Most quality measures are expressed as a rate. The basic construct of a measure is fairly straightforward—numerator, denominator, exclusions, and measure logic. Implementing the measure in a meaningful way requires more precision specified with the appropriate codes sets and/or detailed and precisely-defined data elements.

Figure 10-2 lists key components of the technical specifications used to produce valid and reliable quality measures.

Figure 10-2 Key components of the technical specifications of a quality measure



10.2 Deliverables

- ◆ Updated Measure Information form and Measure Justification form (located in the Measure Development section of Volume 1) to document the updated measure specifications.
- ◆  For contractors developing eMeasures, the Health Quality Measures Format (HQMF) document—which includes a header and a body—should be updated and used in lieu of the MIF. Refer to the eMeasure Specifications During Maintenance Phase for further details on this format.

10.3 Procedure

Step 1: For comprehensive reevaluation or measure updates, update the codes used in the measure

This step does not apply to ad hoc reviews.

Review ICD-10, CPT, LOINC, and any other codes used to identify denominators, numerators, exclusions or exceptions to determine if the underlying code sets have been updated in a way that requires an update to the measure. Document the changes on the MIF in the Release Notes/Summary of Changes section or in a separate document. Refer to the Special Considerations subsection later in this section for timelines on ICD-10 conversion.

Ensure that any new codes reflect the same clinical meaning as their predecessors. If the new codes reflect a change in the clinical meaning compared with the previous codes, clearly document this in the MIF's Release Notes/Summary of Changes section or in a separate document.

Step 2: For comprehensive reevaluations only, review the results of the ongoing environmental scan, the information gathering process, and the audit and validation reports, along with any unsolicited feedback received

When conducting measure updates or ad hoc reviews, this step is not usually necessary. Identify any issues that indicate a problem with the specifications, and propose solutions to the COR/GTL. The Measures Manager may know of similar projects or measures that have experienced similar issues, and may be able to share lessons learned from other contractors' experiences.

When updating technical specifications, alpha or formative testing should be conducted concurrently, as needed, with the development of the new technical specifications. The timing and types of tests performed may vary depending on variables such as data source or complexity of measures. Measures should be specified with the broadest applicability (target population, setting, level of measurement/analysis) as supported by the evidence.²

²National Quality Forum. *Memorandum to the Consensus Standards Approval Committee (CSA)*, from Helen Burstin and Karen Pace; Subject: *Potential Considerations for Quality Performance Measure Construction*. March 7, 2011.



Harmonize with specifications in the NQF-endorsed measures database or QDM when available. For eMeasure development, contractors are expected to use the NQF-developed Measure Authoring Tool and follow the technical specifications guidance in the eMeasure Specifications During Maintenance Phase section.

Step 3: For ad hoc reviews, identify the implications of the issue triggering the review on the technical specifications

An ad hoc review is not an opportunity to fully re-specify a measure. Any changes to the technical specifications during an ad hoc review should focus on the data fields most closely related to the issue triggering the review.

Examples:

- ◆ If the issue is a lack of clarity in the numerator's list of acceptable notations, then the only change should be to that list.
- ◆ If the issue is a missing CPT code from the denominator description, then the only change should be the addition of that code.

Step 4: Review the measure specifications

If any part of the measure is being changed, it is advisable to quickly review the whole measure to ensure the MIF still accurately depicts the measure. For measure updates or ad hoc reviews, this step can usually be done quickly. For comprehensive reevaluations, more attention to detail may be necessary. In either type of maintenance activity, any changes should be clearly documented in the Release Notes/Summary of Changes section of the MIF or in a separate document. The following is a brief guide to the intent of each data field on the MIF.

Review the 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.
- ◆ Long-stay residents with a urinary tract infection.
- ◆ Adults who received 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 help 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

Review 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 4 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.

Review 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 clinical 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.

Review 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.
- ◆ Specification for 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 exclusions or exceptions are still required

If analysis of the measure's performance indicates that the exclusions or exceptions are rarely used or used inappropriately, it may be possible to eliminate them. It is also possible that the ongoing environmental scan or the information gathering process identifies new studies or other literature that provides either rationale for eliminating exclusions/exceptions or an alternative way of handling these cases. If the measure contractor determines that the exclusions or exceptions are still necessary, review them to avoid the possibility of gaming or unintended consequences.

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 picture³. 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.

Define specific exceptions when that structured information can be captured during 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).

³Spertus J A, Bonow RO, Chan 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. Analyzing rates without attention to exception information potentially masks 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).

Exceptions may sometimes be reported as numerator positives instead of being removed from the denominator. Sometimes this is done to preserve denominator size when there is an issue of small numbers. To ensure transparency, capture the allowable exception (either included as numerator positives or removed from the denominator) in a way that can be reported separately, in addition to the overall measure rate.

Denominator exclusions refer to criteria that remove cases 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 exclusions. Missing data may indicate 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 decreases the ability to make valid conclusions about quality.⁴

Example of an exclusion:

- ◆ Patients with bilateral lower extremity amputations who are excluded 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.

⁴The National Quality Forum's *CSAC Guidance on Quality Performance Measure Construction*. May 2011. Available at: http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx, last accessed May 23, 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 During Maintenance Phase section.

Review 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 electronic health records (EHRs). Complete specifications need to be developed and tested for each data source.

Review key terms and data elements

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 EHR, these data elements are defined by the Quality Data Model (QDM). QDM elements should be used when available. When needed quality data elements are not yet available in the QDM, they should be described and presented to NQF to be considered for addition to the QDM. The technical specifications include the “how” and “where” to collect the required data elements. 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 vaccination(s) to be assessed needs 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 During Maintenance Phase section for more

information.) Information that is captured electronically by a medical records system, but is not coded in such a way that a computer program can extract the information, is not considered electronic data because it requires manual abstraction.

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 fee-for-service (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 sufficient information in the details of the denominator, numerator, and exclusions/exceptions, so that each person collecting data for the measure can interpret the specifications in the same manner. If multiple data collection methods are allowed, produce detailed specifications for each method.

Review the level of measurement/analysis

The unit of measurement/analysis is the primary entity upon which the measure is applied. Clearly state and justify the procedure for attributing the measure. Measures should be specified with the broadest applicability (target population, setting, level of measurement/ analysis) as supported by the evidence. However, do not assume that a measure developed for one level is 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, individual physician and physician group.

Review the sampling methodology for clarity

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.

Review the risk adjustment model for clarity

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 for more detailed information.)

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.⁵

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 Measure Information Form or in attachments. Documentation should comply with NQF's open source requirements. 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.

Specifications for risk-adjusted measures should include the following information in the MIF:

- ◆ The method and variables/risk factors (not the details) in the MIF fields.
- ◆ An attachment or URL for the following details: risk model coefficients or equation to estimate each patient's probability for the outcome including coefficients for the variables/risk factors.
- ◆ The codes or definitions for each variable/risk factor.
- ◆ Programming language (e.g., SAS code).

Review any time windows for clarity

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, therefore maintenance may include providing further clarification should that be necessary.

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.

⁵National Quality Forum. *Memorandum to the Consensus Standards Approval Committee (CSA)*, from Helen Burstin and Karen Pace; Subject: *Potential Considerations for Quality Performance Measure Construction*. March 7, 2011.

Review how the measure results are scored and reported for clarity

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 score's interpretation:

- ◆ Better quality = higher score
- ◆ Better quality = lower score
- ◆ Better quality = score within a defined interval
- ◆ Passing score defines better quality

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).⁶

If multiple rates or stratifications are required for reporting, state this in the specifications. If the allowable exclusions are included in the numerator, the measure should be specified to report the overall rate as well as the rate of each allowable exclusion that meets the numerator requirements. Consider 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 to 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.

Review the calculation algorithm for both clarity and accuracy compared to the technical specifications

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

⁶National Quality Forum. *Memorandum to the Consensus Standards Approval Committee (CSA)*, from Helen Burstin and Karen Pace; Subject: *Potential Considerations for Quality Performance Measure Construction*. March 7, 2011.

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 MIF 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 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.

Step 5: Document the measures and obtain the COR/GTL's approval

Review and update the Measure Justification's and the MIF's detailed technical specifications including any additional documents required to produce the measure as it is intended. Clear statements of what has been changed should be documented in the Release Notes/Summary of Changes section of the MIF (a sample is in the Measure Development section) or in a separate document.

Information from measure testing, the public comment period, or other stakeholder input may result in the need to make further changes to the technical specifications. The measure contractor will work with the TEP to incorporate these changes before submitting the measure to the COR/GTL for approval.

The MIF and Measure Justification have been harmonized with the NQF Measure Submission and Endorsement Maintenance forms. By using the MIF to document the technical specifications, environmental scan and testing results, the measure will have the information necessary for NQF's endorsement maintenance activity if CMS decides to do so. If approved by the COR/GTL, an equivalent document that contains the same information/elements as the MIF may be used.

Step 6: Submit the measure to NQF for endorsement maintenance or ad hoc review

NQF conducts three types of measure endorsement maintenance reviews.

- ◆ Annual update—these are usually done annually beginning with year 1 (year 0 is considered the year when endorsement was granted).
- ◆ 3-year endorsement maintenance review—this is done every 3 years beginning with year 3 (year 0 is considered the year when endorsement was granted).
- ◆ Ad hoc endorsement review—this is done only when requested by any interested party or as deemed necessary by NQF.

If the measure developer is contractually required to provide measure maintenance support, the contractor will support CMS during these endorsement maintenance reviews. Measure contractors should follow NQF's online measure submission processes for measure endorsement maintenance, and will provide technical support throughout the NQF endorsement maintenance reviews. This support may include presenting the measure to the steering committee that is evaluating the measure,

or answering questions from NQF staff or the steering committee about the specifications, testing or evidence. The contractor can request a free user account to file the measure submission forms and gain access to the measure dashboard at the following location:

<http://imis.qualityforum.org/Core/CreateAccount.aspx>

Refer to the National Quality Forum Endorsement During Maintenance Phase section for further information. NQF makes periodic updates to the endorsement maintenance process, therefore the NQF Web site should be consulted for the current process:

http://www.qualityforum.org/Measuring_Performance/Maintenance_of_NQF-Endorsed®_Performance_Measures.aspx

During the course of the review, NQF may recommend revisions to the measure. All subsequent revisions must be reported to and approved by the COR/GTL. Update the MIF with any changes agreed upon during the NQF review 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—Centers for Medicare & Medicaid Services (CMS).

10.4 Special Considerations

International Classification of Diseases (ICD)

International Classification of Diseases (ICD) is used for identifying data on claims records, collecting data for use in performance measurement, and reimbursement for Medicare/Medicaid medical claims. ICD is an epidemiological classification code used to identify diagnoses (diseases, injuries, and impairments). The U.S. version also includes procedure codes (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 National Center for Health Statistics (NCHS) director and the CMS administrator. Final decisions are made after the December meeting and become effective October 1 of the following year.⁷

⁷ICD-9-CM Coordination and Maintenance Committee. Available at: http://www.cdc.gov/nchs/icd/icd9cm_maintenance.htm. Accessed August 2, 2012.

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, 2014.⁸

The current system, International Classification of Diseases, 9th Edition, Clinical Modification (ICD-9-CM), does not provide sufficient detail for patients' medical conditions or the procedures and services performed on hospitalized patients.

The new classification system delivers significant improvements through greater information detail 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 in the U.S. for inpatient hospital settings ONLY. The new procedure coding system uses seven alpha or numeric digits while the ICD-9-CM coding system uses three or four numeric digits.⁹

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 were 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, 2014.¹⁰

When a developer submits ICD-10 codes, then the following requirements should also be met:

- ◆ 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.

⁸Centers 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 August 2, 2012.

⁹Centers for Medicare & Medicaid Services. *ICD-10 CM/PCS – An Introduction*. Available at: <http://www.cms.gov/ICD10/Downloads/ICD-10Overview.pdf>. Accessed August 7, 2012.

¹⁰The National Quality Forum, Measure Developer Webinar, June 18, 2012.

- 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 received¹¹

Below is the schedule of updates to ICD-9 and ICD-10 Code Sets during the transition period.

- ◆ October 01, 2011—Last Annual Update to ICD-9 and ICD-10. Code Set Partial Freeze began
- ◆ October 01, 2012—Limited Updates to ICD-9 and ICD-10 for new Technologies
- ◆ October 01, 2013—Claims for services provided on or after this date must use ICD-10 codes for medical diagnosis and inpatient procedures
- ◆ October 01, 2014—Regular updates to ICD-10 code sets begins. Code Set Partial Freeze ends

¹¹The National Quality Forum, *Measure Developer Webinar*, June 18, 2012. Available at:
http://www.qualityforum.org/Calendar/2012/06/Measure_Developer_Webinar.aspx. Accessed: August 2, 2012.

11. eMeasure Specifications During Maintenance Phase

11.1 Introduction

The purpose of this section is to guide measure developers on how to develop, refine 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 maintenance process. (Refer to the Measure Maintenance section for a broader discussion of the types of maintenance.)

In the ongoing environmental scan (refer to the Measure Production and Monitoring section) and the information gathering stage of measure maintenance, the measure contractor identifies issues with the measure that may need to be addressed during the comprehensive reevaluation or an ad hoc review. The contractor may also find other measures with which the CMS measure should be harmonized. The measure contractor should work with its technical expert panel (TEP) to determine the recommended disposition of the measure. Depending on the findings of the information gathering, the TEP will recommend retaining, retiring, revising or suspending the measure. The measure contractor will review those recommendations and make its own recommendation regarding the disposition 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.

Where the measure maintenance process deviates from the maintenance of other types of measures, it is denoted with the following icon (scaled to fit the text):

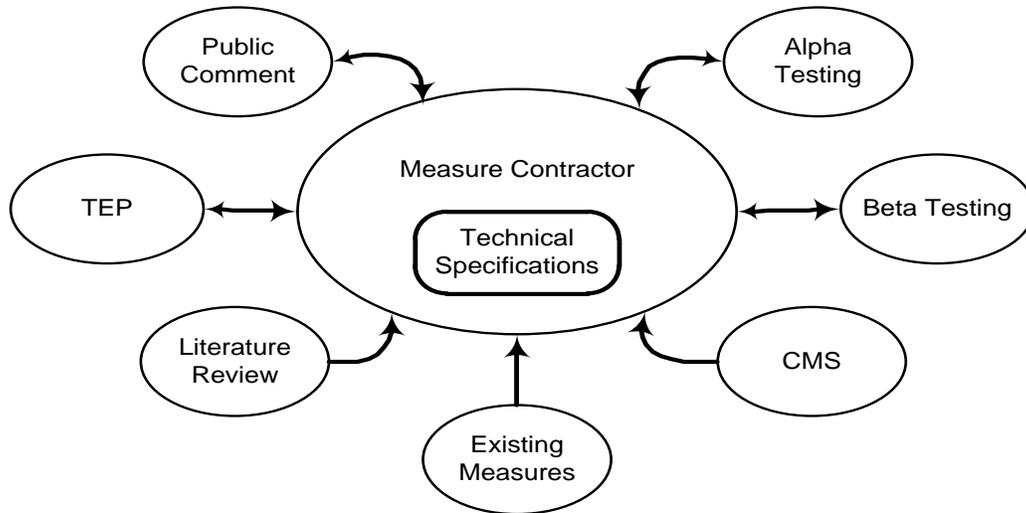
Figure 11-1 eMeasure icon denoting special considerations for eMeasures in the Blueprint.



Developing eMeasure specifications is an iterative process, and as the measure is refined, its specifications should be updated within the National Quality Forum (NQF)-developed Measure Authoring Tool (MAT). If substantial changes are being considered, these changes may require testing of a limited scope. (Refer to the Measure Testing During Maintenance Phase section for an in-depth discussion of testing.) As required during measure development, the measures must be precisely specified and with sufficient details to be distinguishable from other measures and to enable consistent implementation across providers.

Figure 11-2 illustrates the inputs to the process.

Figure 11-2 Factors influencing the development of the measure technical specifications



eMeasures, derived from Clinical Quality Measures for electronic capture, will be authored in the NQF-developed Measure Authoring Tool. This section is a conceptual guide to eMeasure specifications development and maintenance, 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.

As further described in this section, eMeasures are written to conform to the Health Quality Measures Format (HQMF) standard, 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 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 Measure Authoring Tool include the eMeasure XML file, and the eMeasure style sheet.

11.2 Deliverables

- ◆ eMeasure XML file
- ◆ eMeasure style sheet
- ◆ Measure Justification form (required only if the information gathering process results in changes)

11.3 Health Quality Measures Format

A health quality measure (or clinical quality measure) encoded in the Health Quality Measures Format is referred to as an “eMeasure.” 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 query 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 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 Measure Authoring Tool output.

¹<https://mat.qualityforum.org/Login.html>

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 (i.e., 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.

11.4 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) state—context expected for any given QDM element, (4) taxonomy—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 state in 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, *diagnosis*, and the *state* “active” forms the QDM element, *active diabetes diagnosis*, as a specific instance for use in a measure.

11.5 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 category-state pair (e.g., *active diagnosis*) 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 and computer-processable representation of a quality measure.

Thus, at a high level, the process of creating an eMeasure is to map measure data elements to the correct category-state pairs in the QDM, associate each category-state pair with the correct value set(s) to create data criteria, and then assemble the data criteria into population criteria.

²http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx

11.6 Measure Authoring Tool

As described above, NQF has developed a web-based Measure Authoring Tool (MAT), 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 version of the QDM in the January 2012 release of the Measure Authoring Tool is version 2.1.1.1, which is also the version cited under Meaningful Use Stage 2.

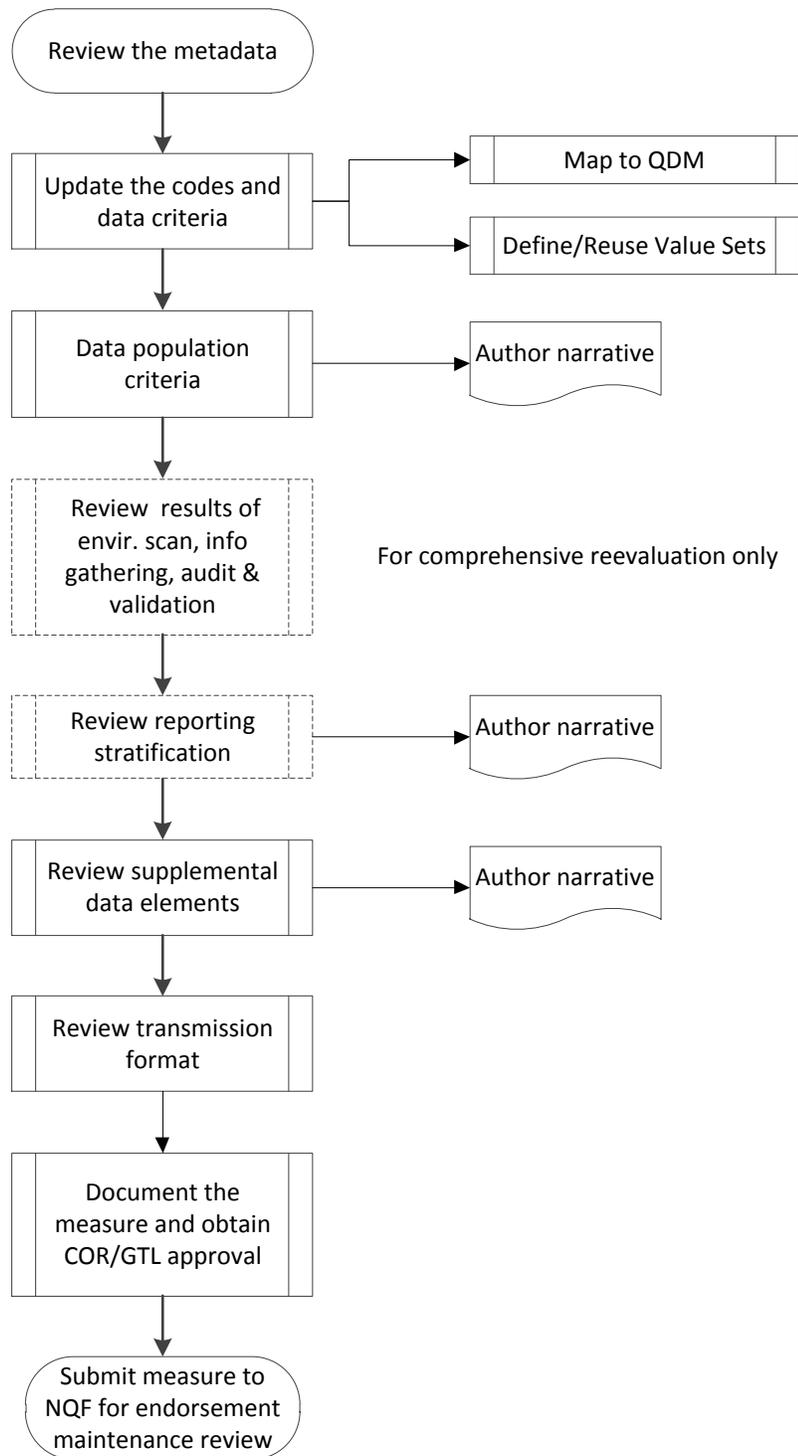
The Measure Authoring Tool does not require measure developers to have an extensive knowledge of the HQMF standard. It is based on the QDM and the building-block approach to creating eMeasures. It supports common use cases and the existing patterns in the pattern library. The Measure Authoring Tool will require ongoing maintenance and support to meet any future measure authoring needs. For instance, if a measure requires a new category-state pair and therefore a new corresponding quality pattern, the pattern must first be developed and then added to the Measure Authoring Tool.

11.7 Procedure

When reviewing an eMeasure for maintenance purposes, the developer should use the process outlined in Figure 11-3 and detailed in the sections that follow. Dashed boxes are only required under certain circumstances.

³http://www.qualityforum.org/Projects/e-g/eMeasure_Format_Review/eMeasure_Format_Review.aspx

Figure 11-3 eMeasure Specifications and Transmission Format Maintenance Process



Step 1: Review the eMeasure 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 11-1 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 11-1. 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 Measure Authoring Tool. “**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 11-1 eMeasure Metadata

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
eMeasure Title	The title of the quality eMeasure.		<u>Required</u>
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.	<u>Required</u>
GUID	Represents the globally unique measure identifier for a particular quality eMeasure.	Field is autopopulated by the Measure Authoring Tool.	<u>Required</u>
eMeasure Version Number	A positive integer value used to indicate the version of the eMeasure.	<p>Displays the integer (whole number) the measure developer enters.</p> <p>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.</p> <p>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.</p>	<u>Required</u>
NQF Number	Specifies the NQF number	<p><i>“Optional” field in MAT</i></p> <p>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.</p>	<u>Not Applicable</u>

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	<u>Required</u>
Measure Steward	The organization responsible for the continued maintenance of the eMeasure.	Can be the same as the Measure Developer.	<u>Required</u>
Measure Developer	The organization that developed the eMeasure.		<u>Required</u>
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).	<u>None</u>
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.”	<u>Required</u>
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”).	<u>None</u>
Disclaimer	Disclaimer information for the eMeasure.	This should be brief.	<u>None</u>
Measure Scoring	Indicates how the calculation is performed for the eMeasure (e.g., proportion, continuous variable, ratio)		<u>Required</u>

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Measure Type	<p>Indicates whether the eMeasure is used to examine a process or an outcome over time</p> <p>(e.g., Structure, Process, Outcome).</p>		<u>Required</u>
Stratification	<p>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</p> <p>(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>)</p>		<u>None</u>
Risk Adjustment	<p>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.</p>	<p>Brief with instructions where complete risk adjustment methodology may be obtained</p>	<u>None</u>

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Rate Aggregation	<p>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.</p> <p>(e.g., pneumonia hospital measures antibiotic selection in the ICU versus non-ICU and then the roll-up of the two).</p>	<p><i>Caution, field should not be used without prior confirmation from eMIG.</i></p>	<p><u>None</u></p>
Rationale	<p>Succinct statement of the need for the measure. Usually includes statements pertaining to Importance criterion: impact, gap in care and evidence.</p>		<p><u>Required</u></p>
Clinical Recommendation Statement	<p>Summary of relevant clinical guidelines or other clinical recommendations supporting this eMeasure.</p>		<p><u>Required</u></p>
Improvement Notation	<p>Information on whether an increase or decrease in score is the preferred result</p> <p>(e.g., a higher score indicates better quality OR a lower score indicates better quality OR quality is within a range).</p>		<p><u>None</u></p>
Measurement Duration	<p>Field will not be used</p>	<p>Does not show in the style sheet.</p>	<p><u>None</u></p>
Reference(s)	<p>Identifies bibliographic citations or references to clinical practice guidelines, sources of evidence, or other relevant materials supporting the intent and rationale of the eMeasure.</p>		<p><u>None</u></p>
Definition	<p>Description of individual terms, provided as needed.</p>	<p>*Note—this field may be removed in the future.</p>	<p><u>None</u></p>

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.		<u>None</u>
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.	<u>None</u>
Initial Patient Population	<p>The <i>initial patient population</i> 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.</p> <p>Details often include information based upon specific age groups, diagnoses, diagnostic and procedure codes, and enrollment periods.</p>	<p>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.</p> <p>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.</p>	<u>Required</u>

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.	<u>Not Applicable</u> (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.)		<u>None</u> (for proportion or ratio eMeasures) <u>Not Applicable</u> (for continuous variable eMeasures)
Numerator	Numerators are <u>used in proportion and ratio eMeasures</u> . 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).		<u>Not Applicable</u> (for continuous variable eMeasures)

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Numerator Exclusions	<p>Numerator Exclusions are <u>used only in ratio eMeasures</u> to define instances that should not be included in the numerator data.</p> <p>(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.)</p>		<p><u>None</u> (for ratio eMeasures)</p> <p><u>Not Applicable</u> (for continuous variable or proportion eMeasures)</p>
Denominator Exceptions	<p>Denominator exceptions are those conditions that should remove a patient, procedure or unit of measurement from the denominator 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 <u>used only in proportion eMeasures</u>. They are not appropriate for ratio or continuous variable eMeasures.</p> <p>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:</p> <ul style="list-style-type: none"> ◆ Medical reasons ◆ Patient reasons ◆ System reasons 	Be specific for medical reasons.	<p><u>None</u> (for proportion eMeasures)</p> <p><u>Not Applicable</u> (for ratio or continuous variable Measures)</p>

Header Data Elements	Definition	Developer Guidance	Preferred Term (Required, None, Not Applicable)
Measure Population	<p>Measure population is used only in continuous variable eMeasures. It is a narrative description of the eMeasure population.</p> <p>(e.g., all patients seen in the Emergency Department during the measurement period).</p>	<p>For continuous variable eMeasures, include the text “Equals All in Initial Patient Population.” Then add any specific additional criteria if needed.</p>	<p><u>Not Applicable</u></p> <p>(for ratio or proportion eMeasures)</p>
Measure Observations	<p>Measure observations are used only in continuous variable eMeasures. They provide the description of how to evaluate performance,</p> <p>(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.</p>		<p><u>Not Applicable</u></p> <p>(for ratio or proportion eMeasures)</p>
Supplemental Data Elements	<p>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.</p> <p>Additional supplemental data elements required for risk adjustment or other purposes of data aggregation can be included in the Supplemental Data Element section.</p>	<p>Due to the four CMS required fields, the Developer must always populate with payer, ethnicity, race and sex.</p> <p>For CMS measures use the following language in Supplemental Data section (January 2012 release of MAT)</p> <p>“For every patient evaluated by this measure also identify payer, race, ethnicity and sex.”</p> <p>Other information may be added for other measures.</p>	<p><u>Required</u></p>

Conventions for developing eMeasures

Developers should use the following conventions in Table 11-2 when developing the CMS eMeasures to ensure standardization of the measures for display of specifications and quality reporting:

Table 11-2 Conventions for Developing eMeasures

Data Element or Item	Convention
Calculation of Age for Ambulatory Measures	<p>To be included in the eMeasure, the patient’s age must be \geq IPP inclusion criterion before the start of the Measurement Period. The Measurement Period is January 1-December 31, 20xx.</p> <p>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.</p>
Calculation of Age for Hospital Measures	<p>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:</p> <ol style="list-style-type: none"> 1. Patient age at admission (in years) [refer to example in note A below] = Admission date – Birth date 2. Newborn patient age at admission (in days) [refer to example in note B below] = Admission date - Birth date 3. Patient age at discharge (in years) [refer to example in note C below] = Discharge date - Birth date 4. Patient age at time of event (in years) [refer to example in note D below] = Event date - Birth date <p>Note A—An example of a corresponding representation in HQMF for an IPP inclusion criterion for patient age range:</p> <ul style="list-style-type: none"> • "Patient Characteristic Birthdate: birth date" \leq 65 years starts before start of x; AND • "Patient Characteristic Birthdate: birth date" \geq 18 years starts before start of x <p>Note B—An example of a corresponding representation in HQMF: Patient is \leq 20 days old at time of admission:</p> <ul style="list-style-type: none"> • "Patient Characteristic Birthdate: birth date" \leq 20 days starts before start of "Encounter Performed: hospital encounter (admission)" <p>Note C—An example of a corresponding representation in HQMF: Patient is \geq 10 years old at time of discharge:</p> <ul style="list-style-type: none"> • "Patient Characteristic Birthdate: birth date" \geq 10 years starts before start of "Encounter Performed: hospital encounter (discharge)" <p>Note D—An example of corresponding representation in HQMF: Patient is \leq 21 at time of hepatitis vaccination:</p> <ul style="list-style-type: none"> • "Patient Characteristic Birthdate: birth date" \leq 21 years starts before start of "Medication Administered: hepatitis vaccine (date time)"

Step 2: Update the codes used in the measure

This step does not apply to ad hoc reviews.

Review vocabularies and value sets used to identify denominators, numerators, exclusions or exceptions to determine if the underlying code sets have been updated in a way that requires an update to the measure. Ensure that any new codes reflect the same clinical meaning as their predecessors. If the new codes reflect a change in the clinical meaning compared with the previous codes, clearly document the changes in a separate document (e.g., Release Notes/Summary of Changes document).

Map to QDM

As noted above, the process of creating an eMeasure is to map measure data elements to the correct category-state pairs in the QDM, associate each category-state pair 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 vs. new measures.

Retooled measures

Measure developers need to identify data elements in the existing paper-based measure and map them to QDM category-state pair 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 category-state pair 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 category-state pair. Developers will associate each QDM category-state pair with an existing value set or create a new value set if one does not exist, and define additional attributes if applicable.

Review any time windows for clarity

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, therefore maintenance may include providing further clarification should that be necessary.

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.

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 11-3 through 11-6 below. Tables 11-5 and 11-6 are complementary to one another, with the same general information being provided, but from two different starting points: 11-5 illustrates the vocabularies as they relate to the clinical concepts of the QDM, while 11-6 is from the perspective of the clinical concepts, according to the QDM, but with respect to the appropriate vocabularies. Measures that are retooled or developed now should 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 11-3 ONC HIT Standards Committee Recommended Vocabulary Standards Summary

Clinical Vocabulary Standards:	Others:
SNOMED CT	CVX
LOINC	CDC-PHIN/VADS
RxNorm	UCUM
	ISO-639

⁴Centers 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.

Table 11-4 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
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 11-5 Quality Data Model Categories with ONC HIT Standards Committee Recommended Vocabularies

Quality Data Model Category or Clinical Concept	Quality Data Model Type (State)	Clinical Vocabulary Standards	Transition Vocabulary
Adverse Effect	Allergy: Reaction (Documented, Updated)	SNOMED CT	N/A
	Allergy: (Documented, Updated) Attribute: Causative agent: Medication [The medication agent to which the patient is allergic – the causative agent.]	RxNorm	
	Allergy: (Documented, Updated) Attribute: Causative agent: Non-medication Substance [The non-medication agent to which the patient is allergic – the causative agent.]	SNOMED CT	
	Non-Allergy: Reaction (Documented, Updated)	SNOMED CT	
	Non-Allergy: (Documented, Updated) Attribute: Causative agent: Medication [The medication agent to which the patient is allergic – the causative agent.]	RxNorm	
	Non-Allergy: (Documented, Updated) Attribute: Causative agent: Non-medication Substance [The non-medication agent to which the patient is allergic – the causative agent.]	SNOMED CT	
Non-medication substance	Substance (Administered, Ordered, Documented) [Substance can be an attribute of an adverse effect—the causative agent, or it can be a stand-alone element, e.g., Substance administered: enteral supplements]	SNOMED CT	N/A
Condition; Diagnosis; Problem, family history	Condition/Diagnosis/Problem (Active, Inactive, Resolved)	SNOMED CT	ICD-9-CM, ICD-10-CM
Symptom	Symptom (Active, Assess, Inactive, Resolved)	SNOMED CT	N/A
Family history	Family History (Reported, Updated, Declined)	SNOMED CT-	ICD-9-CM, ICD-10-CM

Quality Data Model Category or Clinical Concept	Quality Data Model Type (State)	Clinical Vocabulary Standards	Transition Vocabulary
Encounter (any patient-provider interaction, e.g. phone call, email regardless of reimbursement status, status—includes traditional face-to-face encounters)	Encounter (also referred to as Patient-Provider Interaction) (Ordered, Performed, Recommended, Declined)	SNOMED CT	CPT, HCPCS, ICD-9-CM Procedures, ICD-10-PCS
Device	Device (Applied, Ordered, Planned, Declined)	SNOMED CT	N/A
Physical exam finding	Physical Exam (Ordered, Performed, Recommended, Declined)	SNOMED CT	N/A
Laboratory test names	Laboratory Test (Ordered, Performed, Declined)	LOINC	N/A
Laboratory test results	Laboratory Test (Ordered, Performed, Declined)	SNOMED CT	N/A
	Units of Measure	UCUM-(The Unified Code for Units of Measure)	
Diagnostic study test names	Diagnostic Study (Ordered, Performed, Recommended, Declined)	LOINC	HCPCS
Diagnostic study test results	Diagnostic Study (Ordered, Performed, Recommended, Declined)	SNOMED CT	N/A
	Units of Measure	UCUM-(The Unified Code for Units of Measure)	
Intervention (Note: Intervention is being retired—it is incorporated under Procedure)	Intervention (Ordered, Performed, Recommended, Declined)	SNOMED CT	CPT, HCPCS, ICD-9-CM Procedures, ICD-10-PCS
Procedure	Procedure (Ordered, Performed, Recommended, Declined)	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)	Characteristic (also referred to as Individual Characteristic or Patient Characteristic) (Documented, Declined)	SNOMED CT	N/A
Functional Status	Functional Status (Ordered, Performed, Declined)	SNOMED CT	N/A

Quality Data Model Category or Clinical Concept	Quality Data Model Type (State)	Clinical Vocabulary Standards	Transition Vocabulary
Categories of function	Functional Status (Ordered, Performed, Declined)	ICF-(International Classification of Functioning, Disability, and Health)	N/A
Communication	Communication (Acknowledge, Decline, Record, Transmit)	SNOMED CT	CPT, HCPCS
Assessment instrument questions (questions for patient preference, experience, characteristics)	Characteristic (also referred to as Individual Characteristic or Patient Characteristic) (Documented, Declined)	LOINC	N/A
Medications (administered, excluding vaccines)	Medication (Active, Administered, Order, Dispense, Decline)	RxNorm	N/A
Vaccines (administered)	Medication (Active, Administered, Order, Dispense, Decline)	CVX	N/A
Patient characteristic (Administrative Gender, DOB)	Characteristic (also referred to as Individual Characteristic or Patient Characteristic) (Documented, Declined)	CDC-Public Health Information Network (PHIN)/Vocabulary Access and Distribution System (VADS) http://www.cdc.gov/phin/activities/vocabulary.html	N/A
Patient Characteristic (Ethnicity, Race)	Characteristic (also referred to as Individual Characteristic or Patient Characteristic) (Documented, Declined)	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)	Characteristic (also referred to as Individual Characteristic or Patient Characteristic) (Documented, Declined)	ISO-639-1:2002	N/A

Quality Data Model Category or Clinical Concept	Quality Data Model Type (State)	Clinical Vocabulary Standards	Transition Vocabulary
Patient characteristic (Payer)	Characteristic (also referred to as Individual Characteristic or Patient Characteristic) (Documented, Declined)	<p>Payer Typology (Public Health Data Standards Consortium Payer Typology) (scope)http://www.phdsc.org/standards/pdfs/SourceofPaymentTypologyVersion5.0.pdf</p> <p>In PHIN VADS (vocabulary) http://phinvads.cdc.gov/vads/ViewCodeSystemConcept.action?oid=2.16.840.1.113883.221</p>	N/A

Table 11-6 ONC HIT Standards Committee Recommended Vocabulary Standards with Quality Data Model Categories

Vocabulary	General Clinical Concept	QDM Category (October 2011+)
SNOMED CT	Allergies: Non-medication substance [The non-medication agent to which the patient is allergic – the causative agent.]	Adverse Effect: Allergy (causative agent)
	Non-allergic adverse effects (e.g., intolerance)	Adverse Effect: Non-allergy (causative agent)
	Non-medication substances (e.g., latex) [Substance can be an attribute of an adverse effect—the causative agent, or it can be a stand-alone element, e.g., Substance administered: enteral supplements]	Substance (Substance administered: enteral supplements; Adverse effect: allergy (causative agent: latex)
	Artifacts of communication (e.g., medicine list, clinical summary)	Communication
	Disorders	Condition, Diagnosis, Problem

Vocabulary	General Clinical Concept	QDM Category (October 2011+)
	Diseases	
	Conditions	
	Problems	
	Symptoms (e.g., nausea, vomiting, pain [reported by patient])	Symptom
	Patient provider interaction (any type; e.g., phone calls, etc.; regardless of reimbursement status— includes traditional face-to-face encounters)	Encounter <i>(also referred to as Patient-Provider Interaction)</i>
	Instruments, Hardware	Device
	Results and findings for: <ul style="list-style-type: none"> ▪ Laboratory results ▪ Diagnostic studies ▪ Physical exam 	Physical Exam
		Laboratory Exam
		Diagnostic Study (non-laboratory)
	Procedures: <ul style="list-style-type: none"> ▪ Surgical ▪ Physical Manipulation ▪ Counseling ▪ Education Results and findings for procedures	Procedure
	Expected answers to questions about: <ul style="list-style-type: none"> ▪ Patient characteristics ▪ Patient experience ▪ Patient preference ▪ Risk evaluation ▪ Family history Functional status (e.g., answers to assessment instruments such as “patient has a caregiver”)	Characteristics <i>(also referred to as Individual or Patient Characteristic)</i>
		Experience <i>(also referred to as Individual or Patient Experience)</i>
		Preference <i>(also referred to as Individual or Patient Preference)</i>
Risk Evaluation		
Family History		
Functional Status		
LOINC	Assessment Instruments Assessment Questions	Functional Status
		Characteristics
		Experience
		Preference
		Risk Evaluation

Vocabulary	General Clinical Concept	QDM Category (October 2011+)
		Family History
	Laboratory test names	Laboratory Test
	Diagnostic study names	Diagnostic study (non-laboratory)
	Staffing Resources (e.g., Nursing units)	System Resources
UCUM (The Unified Code for Units of Measure)	Units of measure	Diagnostic Study Results
		Laboratory Test Results
RxNorm	Medications causing allergic reactions	Adverse Effect: Allergy
	Medications causing non-allergic adverse effects (e.g., intolerance)	Adverse Effect: Non-Allergy
	Medications administered (excluding vaccines)	Medication
Payer Typology (Public Health Data Standards Consortium Payer Typology)	Payer	Characteristics <i>(also referred to as Individual or Patient Characteristic)</i>
OMB Ethnicity/Race Value Set	Ethnicity, Race	Characteristics <i>(also referred to as Individual or Patient Characteristic)</i>
ISO 639-1:2002	Preferred Language	Characteristics <i>(also referred to as Individual or Patient Characteristic)</i>
ICF (International Classification of Functioning, Disability, and Health)	Categories of Function	Functional Status
CVX	Vaccines administered	Medication
CDC-PHIN/VADS (Public Health Information Network/Vocabulary Access and Distribution System)	Administrative Gender, DOB	Characteristics <i>(also referred to as Individual or Patient Characteristic)</i>

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 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 Value Set Authority Center established by National Library of Medicine (NLM)⁷. The Value Set Authority Center is a publicly available authoritative repository of controlled value sets in which NLM will support ongoing maintenance.

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

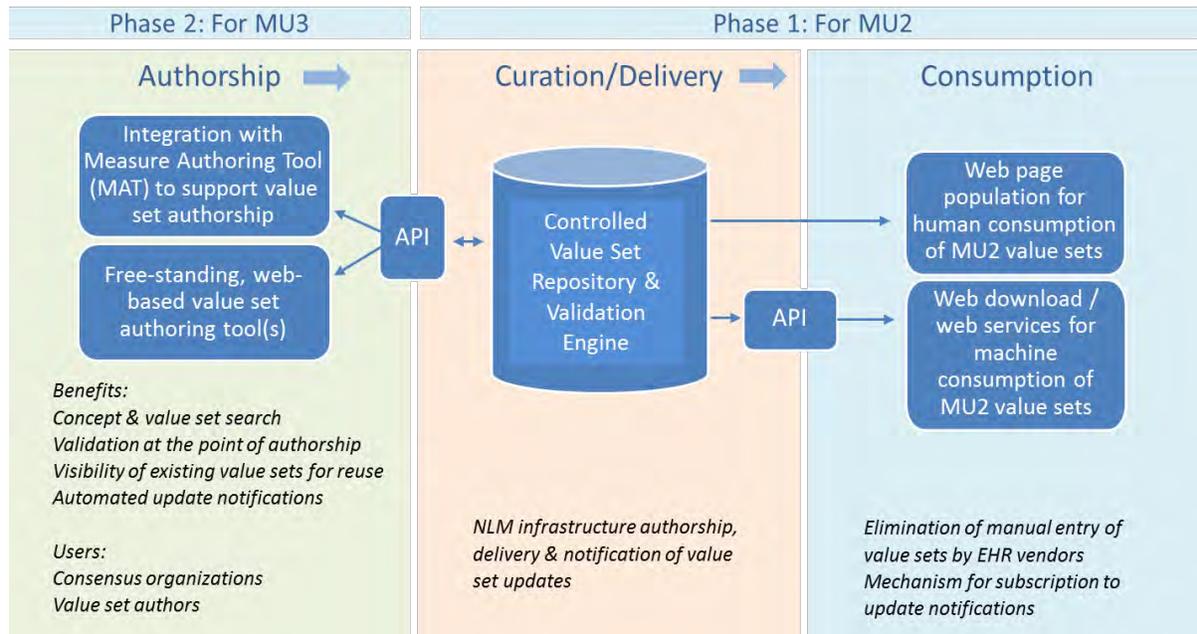
⁵U.S. SNOMED CT Content Request System. <https://uscrs.nlm.nih.gov/>

⁶UMLS Terminology Services. <https://uts.nlm.nih.gov//home.html>

⁷Value Set Authority Center, National Library of Medicine. <http://vsac.nlm.nih.gov>

integrated into measure authoring tools to support value set development. Figure 11-4 shows the National Library of Medicine’s vision for value set management.

Figure 11-4 Vision for Robust Value Set Management⁸



For the value set quality assurance checks, NLM assesses 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 make the corrections to the value sets in the Measure Authoring Tool.

Step 3: Review 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 the Table 11-7. The definitions for these populations are defined in Appendix 11d.

⁸U.S. National Library of Medicine: Value Set Validation: Author Info and Instructions.

Table 11-7 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
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 11c.

Determine if denominator exclusions or exceptions are still required

If the analysis of the measure’s performance indicates that the exclusions or exceptions are rarely used or used inappropriately, it may be possible to eliminate them. It is also possible that the ongoing environmental scan or the information gathering process identifies new studies or other literature that provides either rationale for eliminating exclusions/exceptions or an alternative way of handling these cases. If the measure contractor determines that the exclusions or exceptions are still necessary, review them to avoid the possibility of gaming or unintended consequences.

Definitions of each population are presented in the definitions in Table 11-1. Refer to the Technical Specifications During Maintenance Phase 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 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 Measure Authoring Tool documentation for adding these relationships into an eMeasure.

EHR Context

“EHR Context” defines where to look in an EHR to find the data needed to resolve a criterion. For instance, if the EHR 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 EHR Contexts in

⁹http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx#t=2&s=&p=4%7C

eMeasures are more suggestive than prescriptive. In some cases, particularly where Meaningful Use has not yet standardized the context (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 EHRs. Consistent application of a standard set of contexts will lower the barrier to automated electronic reporting of quality and thus lead to consistent data storage in EHRs in the future.

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 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 4: For comprehensive reevaluations only, review the results of the ongoing environmental scan, the information gathering process, and the audit and validation reports, along with any unsolicited feedback received

When conducting measure updates or ad hoc reviews, this step is not usually necessary. Identify any issues that indicate a problem with the specifications, and propose solutions to the COR/GTL. The Measures Manager may know of similar projects or measures that have experienced similar issues, and may be able to share lessons learned from other contractors’ experiences.

When updating technical specifications, alpha or formative testing should be conducted concurrently, as needed, with the development of the new technical specifications. The timing and types of tests performed may vary depending on variables such as data source or complexity of measures. Measures should be specified with the broadest applicability (target population, setting, level of measurement/analysis) as supported by the evidence.¹⁰

Step 5: Review 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.

¹⁰National Quality Forum. Memorandum to the Consensus Standards Approval Committee (CSA), from Helen Burstin and Karen Pace; Subject: Potential Considerations for Quality Performance Measure Construction. March 7, 2011.

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 6: Review the 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. Value sets used to define the supplemental data elements are subject to change, and therefore the measure should be reviewed to ensure that the appropriate ones are being used.

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 gender 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 Set¹¹ (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 Set¹² (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

¹¹The 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).

¹²The 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).

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 7: Review 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 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 III¹³, or Physician Quality Reporting Initiative (PQRI) XML format¹⁴. The measure developer should consult with their COR/GTL to determine what transmission format should be used. For example, your 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.

¹³QRDA 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.

¹⁴http://www.cms.gov/PQRS/Downloads/PQRI_2010RegistryXMLSpecifications.pdf

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 11-5.

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 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 11-5 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
<ul style="list-style-type: none"> • Measure Section • Reporting Parameters • Patient Data

Measure Section
<ul style="list-style-type: none"> • Measure: Surgery Patients with Appropriate Hair Removal (NQF0301)

Reporting Parameters
<ul style="list-style-type: none"> • Reporting period: 01 Jan 2012 - 31 Dec 2012

Patient Data
<ul style="list-style-type: none"> • 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 specification¹⁵ 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.

Step 8: Document the measure and obtain COR/GTL approval

Development of the complete technical specifications is an iterative process including input from the original measure steward (for retooled measures), information gathering, TEP, public comments and measure testing. (Refer to the Measure Testing During Maintenance Phase section for guidance on alpha and beta testing methods during measure development). The measure contractor should complete the detailed technical specifications and any additional documents required to produce the measure as it is intended (e.g., risk adjustment methodologies, etc.), and work with the TEP to incorporate these changes before submitting the measure to the COR/GTL for approval.

The complete eMeasure specifications are documented in the NQF-developed Measure Authoring Tool (for both retooled and de novo measures) and Measure Justification form (for de novo measures only). Clear statements of what has been changed within the measure should be acknowledged in a separate document (e.g., a Release Notes/Summary of Changes document).

Step 9: Submit the eMeasure to NQF for endorsement maintenance or ad hoc review

NQF conducts three types of measure endorsement maintenance reviews.

- ◆ Annual update—these are usually done annually beginning with year 1 (year 0 is considered the year when endorsement was granted).
- ◆ 3-year endorsement maintenance review—this is done every 3 years beginning with year 3 (year 0 is considered the year when endorsement was granted).

¹⁵http://www.cms.gov/PQRS/Downloads/PQRI_2010RegistryXMLSpecifications.pdf

- ◆ Ad hoc endorsement review—this is done only when requested by any interested party or as deemed necessary by NQF.

If the measure developer is contractually required to provide measure maintenance support, the contractor will support CMS during these endorsement maintenance reviews. Measure contractors should follow NQF's online measure submission processes for measure endorsement maintenance, and will provide technical support throughout the NQF endorsement maintenance reviews. This support may include presenting the measure to the steering committee that is evaluating the measure, or answering questions from NQF staff or the steering committee about the specifications, testing or evidence. eMeasure developers should also refer to the eMeasure considerations within the Measure Testing During Maintenance Phase section for further information on testing requirements during maintenance.

Developers can request a free user account to file the measure submission forms and gain access to the measure dashboard at the following location:

<http://imis.qualityforum.org/Core/CreateAccount.aspx>

Refer to the National Quality Forum Endorsement Maintenance During Maintenance Phase section for further information. NQF makes periodic updates to the endorsement maintenance process, therefore the NQF Web site should be consulted for the current process:

http://www.qualityforum.org/Measuring_Performance/Maintenance_of_NQF-Endorsed®_Performance_Measures.aspx

During the course of the review, NQF may recommend revisions to the measure. All subsequent revisions must be reported to and approved by the COR/GTL. Update the eMeasure with any changes agreed upon during the NQF review 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—Centers for Medicare & Medicaid Services (CMS).

11.8 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.

Appendix 11a¹⁶ XML View of a Sample eMeasure¹⁷

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		

¹⁶Note that while 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.

¹⁷This is a sample measure, solely for illustration purposes. It is not a real measure.

Measure scoring	Proportion
Measure type	Process
Stratification	<p>Population 1: DOB 14-23 years before start of measurement period</p> <p>Population 2: DOB 14-19 years before start of measurement period</p> <p>Population 3: DOB 20-23 years before start of measurement period</p>
Risk Adjustment	None
Rate Aggregation	None
Rationale	<p>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.</p>
Clinical Recommendation Statement	<p>To determine whether the goals of therapy are being met, monitoring is recommended in the 6 areas listed below:</p> <ul style="list-style-type: none"> ◆ 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	<p>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.</p>

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 criteria¹⁸

- ◆ 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 =

¹⁸ Populations will vary based on type of measure scoring.

- AND:
 - 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 11b 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>
...
```

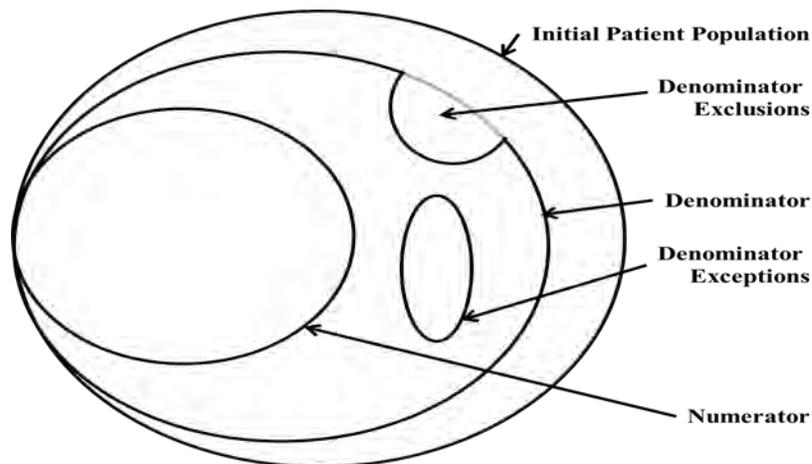
```
</section>
<section>
<title>Reporting stratification</title>
<text>... narrative reporting stratification descriptions ...</text>
<entry>... Formal reporting stratification definition ...</entry>
<entry>... Formal reporting stratification definition ...</entry>
...
</section>
<section>
<title>Supplemental 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 11c 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 11-1 (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.¹⁹

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 criteria.

¹⁹Many 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:

$$\text{Performance Rate} = (\text{NUMER}) / (\text{DENOM} - \text{EXCL} - \text{EXCEP})$$

From the example above, counting all individuals within the population, the following aggregate counts are determined:

Initial Patient Population: N=150 (i.e. 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).

$$\text{Performance Rate} = (\text{NUMER}) / (\text{DENOM} - \text{EXCL} - \text{EXCEP}) = (75)/(150-20-5) = 0.6$$

Appendix 11d Definitions

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 11e Acronyms and Abbreviations

ANSI	American National Standards Institute
ASTM	ASTM International, originally known as the <i>American Society for Testing and Materials</i>
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

12. Technical Expert Panels During Maintenance Phase

12.1 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. Consulting with the TEP is a very important in the comprehensive reevaluation process because it ensures transparency and allows an opportunity to obtain balanced, multi-stakeholder input. A TEP may also be consulted during an ad hoc review of the measure. The National Quality Forum (NQF) or the measure steward may request ad hoc reviews at any time. These requests may come as a result of concerns that evidence supporting the measure has changed or unintended consequences of measure results.

For most measure maintenance contracts, the measure developer will convene several TEP meetings, either by teleconference or in person. The TEP will be asked to review and comment on the literature review, guideline changes, performance data, and related measures. They may be asked to reevaluate the measure in light of other measures that might serve as replacements or as candidates for harmonization. The members may also review the measure specifications, current evidence and data supporting the measure and provide their input regarding any modifications to the measure specification, scientific soundness and feasibility of the measure.

The contractor should keep an overall vision for discerning the breadth of quality concerns and related goals for improvement identified for the setting of care. The TEP should be directed and encouraged to think broadly about principal areas of concern regarding quality as they are related to the topic or contract at hand. Finally, at the end of the comprehensive reevaluation process, the contractor should be able to show how the recommended measure continue to relate to overall Department of Health and Human Services (HHS) goals including the National Quality Strategy priorities, the Centers for Medicare & Medicaid Services (CMS) programs, and recommendations of the Measure Application Partnership.

Those TEP members who were involved in the measure development provide an excellent foundation for the maintenance TEP because

- ◆ They are already familiar with the measure's strengths.
- ◆ They know where there may be an opportunity to improve the measure.
- ◆ They understand the measure's evolution from concept to evaluation.

If TEP members involved during measure development are not available, and a new TEP or recruiting additional TEP members is necessary, then begin the process of posting the call for nominations 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 at the beginning of the comprehensive reevaluation.

The TEP should include recognized experts in relevant fields, including: 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.

12.2 Deliverables

- ◆ Call for TEP, if necessary
- ◆ TEP Nomination/Disclosure/Agreement forms for any new TEP members
- ◆ TEP Charter, if revision is needed
- ◆ TEP membership list
- ◆ Meeting minutes
- ◆ Measures presented to the TEP for reevaluation and any new measures that the TEP recommends as replacement or to augment existing measures
- ◆ Measure Evaluation reports
- ◆ Updated Measure Information Form and Measure Justification form, if they are modified after the TEP meetings
- ◆ TEP Summary report

12.3 Procedure

The following steps should be performed when convening a TEP and conducting the TEP meetings. Updates to the Measure Information Form and the Measure Justification form may be made after the TEP deliberations.

Some measures already have TEPs overseeing their maintenance. In these cases, it may not be necessary to recruit new members. However, recruitment may be needed in the case of attrition, if the measure contractor identifies a new stakeholder community that needs to be represented or if the measure contractor wishes to broaden representation.

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). Measure contractors will work with the Measures Manager to achieve these postings. Web site postings process 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 maintenance project
- ◆ Overall vision for discerning the breadth of quality concerns and related goals for improvement identified for the setting of care
- ◆ Specific objectives
- ◆ Measure maintenance processes

- ◆ Types of expertise needed (topic/subject expert, performance measurement, quality improvement, consumer perspective, purchaser perspective, health care disparities, etc.)
- ◆ Expected time commitment and anticipated meeting dates and locations, including any ongoing involvement that is expected to occur throughout the measure maintenance review process
- ◆ Instructions for required information (TEP Nomination/Disclosure form, letter of intent, etc.)
- ◆ Contractor's email address where TEP nominations and any questions are to be sent

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 Contracting Officer Representative/Government Task Leader (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:

- ◆ 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.
- ◆ 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:

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

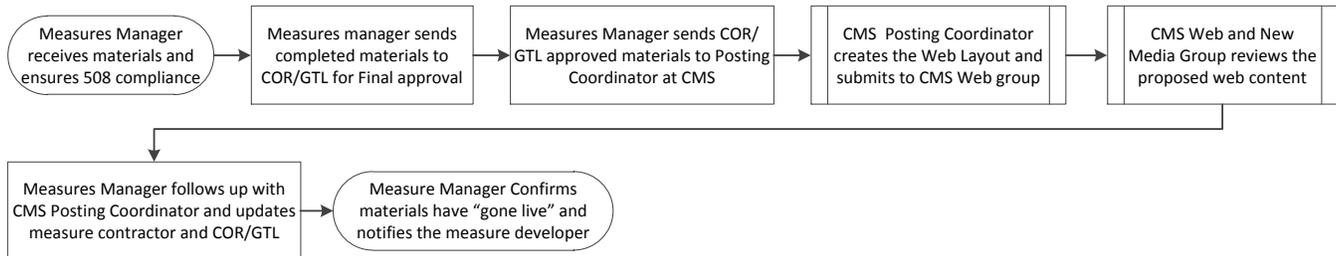
Step 3: Post the call for nominations

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 12-1 The Posting Process



The posting process:

1. After receiving the materials, the Measures Manager reviews them to confirm they are Section 508 compliant. Information about CMS 508 Compliance is available at <http://www.hhs.gov/web/508/index.html>.
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.

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.

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.

Section 508 (212 U.S.C. § 7124d) requires federal agencies to provide employees and members of the public with disabilities access to electronic and information technology that is comparable to the access available to individuals without disabilities. The law applies to all federal agencies when they develop, procure, maintain, or use electronic and information technology (EIT). Therefore, the measure contractor must ensure that all documents posted on the CMS Web site are 508 compliant. Information about CMS 508 Compliance is available at <http://www.hhs.gov/web/508/index.html>.

Step 4: Inform the COR/GTL about the TEP members

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. The following factors should be considered in proposing and selecting the 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 120 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—select individuals who will not be inappropriately influenced by any special interest. TEP members are asked to disclose any potential conflict of interest during the nomination process. 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 the COR/GTL to decide if the individual’s interest or relationships may affect the discussions or conclusions.

Document the proposed TEP member’s name and credentials, organizational affiliation, city and state, area of expertise and experience. Include brief points to clearly indicate why the 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 elected TEP chair and co-chair members 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 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. The Measures Manager will provide examples upon request. The information should be available until the TEP Summary report is removed from the Web site.

Step 8: Write/Review 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)

In the case of a pre-existing TEP, the charter should be reviewed to ensure it includes the maintenance activities that the TEP is now expected to perform. Update it as necessary. 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. E-mail 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.

Measure contractor may decide that additional experts and staff may be needed to support the TEP. The areas of expertise may include, but not limited to, data management and coding experts, electronic medical records experts, health informatics experts, 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 measure reevaluation, including the measure evaluation criteria.
- ◆ Documentation of the measure. Depending on the number of measures that the TEP will review, the contractor may modify or shorten the Measure Information Form and 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.
- ◆ Other documents as applicable.

Step 11: Conduct the TEP meetings

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.

During the first meeting, review and ratify the TEP Charter, explaining the TEP's role and scope of responsibilities. Present the findings of the literature review and environmental scan any new measures which might be a competing measure with the measure under review for maintenance. Discuss any overall quality concerns, such as measurement gaps and alignment across programs and settings as well as overarching goals for improvement.

Keep detailed minutes of all TEP meetings. 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 will 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 12: Reevaluate the measure for maintenance

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. Refer to the Measure Evaluation During Maintenance Phase section for detailed instructions.

Measure contractor may conduct a preliminary evaluation of the measure and complete a draft Measure Evaluation report before the TEP meeting. These drafts can be presented to the TEP for discussion. Measure contractor may use the TEP discussions as input to complete the Measure Evaluation report for the measure after the meeting. 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.

Step 13: Prepare the TEP summary report

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 reevaluated measures meet the overall quality concerns and goals for improvement
- ◆ Dates of meetings
- ◆ TEP composition
- ◆ Recommendations on the reevaluated measures

Step 14: Post the summary report

Work with the Measures Manager to post the approved TEP Summary report. 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 During Maintenance Phase section.

Template for Call for TEP

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/112_MeasuresManagementSystemBlueprint.asp#TopOfPage.

TEP requirements:

Complete the following if applicable: <A TEP currently exists for <project name>. We are recruiting for additional members to <describe the need for additional members (area of expertise, etc.)>.

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
- ◆ 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>

If they are still available, the TEP members involved in measure development provide an excellent foundation for the maintenance TEP. If this is the case, list current TEP members already on the panel:

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 e-mail to: **<insert e-mail 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. Conflict of interest glossary of terms can be found at https://www.cms.gov/MMS/15_TechnicalExpertPanels.asp#TopOfPage.

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):

E-mail 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):

E-mail 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 120 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: _____

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:

13. Public Comment During Maintenance Phase

13.1 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 being reevaluated and can provide critical suggestions not previously considered by the measure contractor or its technical expert panel (TEP).

Measures undergoing a comprehensive reevaluation may require a public comment period. However, if the measures in question fall within the rulemaking processes (e.g., Inpatient Prospective Payment System or Physician Fee Schedule), the public comment solicited through the rulemaking process satisfies this step. A public comment period can also be useful for ad hoc reviews that result in significant changes to the measure.

The results of the comprehensive reevaluation should be posted on the dedicated CMS Web site for public comment if the results meet any of the following criteria:

- ◆ Material change in the measure
- ◆ When results of the comprehensive reevaluation will not be included in the proposed rule
- ◆ As directed by Contracting Officer Representative/Government Task Leader (COR/GTL)

13.2 Deliverables

- ◆ Call for Public Comment
- ◆ List of Stakeholders
- ◆ List of the measures with MIFs and Measure Justification forms
- ◆ Verbatim Public Comments
- ◆ Public Comment Summary report

13.3 Procedure

The following steps are essential to soliciting public comment. Deviation from the following procedure requires the COR/GTL's approval.

The call for public comment involves 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 achieve these postings. Web site postings involve two CMS divisions, and the process to post each set of materials will take approximately five working days.

Step 1: Determine need for solicitation of public comment

When the measures are undergoing comprehensive reevaluation, the results of the review may determine that a public comment period is needed. Measure contractors should solicit public comment if any of the criteria described in the introduction of this section are met.

Step 2: Write the call for public comment

Prepare the Call for Public Comment during Measure Maintenance. Measure contractors may use this document for designing 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.

Step 3: 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 the contractor obtains the approval, notify the stakeholder organizations before the posting goes live or simultaneously with the posting on the Web site. Relevant stakeholder groups may include, but are not limited to:

- ◆ Quality alliances (AQA, 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.
- ◆ 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 Listserv mailing list.
- ◆ Use social media (e.g., Twitter, Facebook, YouTube, LinkedIn, etc.) See your COR/GTL for the process.

Step 3: Post the measures following the 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.”

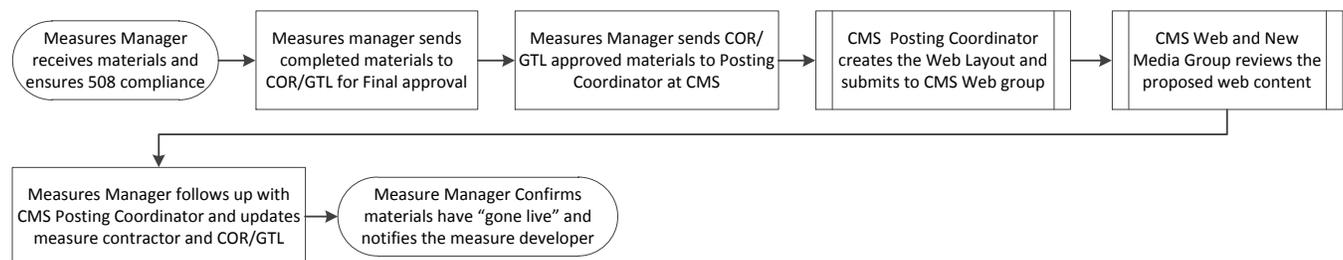
As a general rule, the call should be posted on the Web site for at least two weeks to allow sufficient time for the public to provide comments. The COR/GTL will make the final decision as to how long the call should be posted.

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 maintenance contract.
- ◆ The processes used to review the measures, for example:
 - Identifying important quality goals related to Medicare services.
 - Conducting literature reviews and grading evidence.
 - Refining the specifications for each quality measure based on feedback from stakeholders, performance, guideline changes, and the literature reviews.
 - Analyzing performance trends of the measures and sub-group/disparities analyses.
 - Obtaining reevaluation of proposed measures by technical expert panels. (As directed by the COR/GTL, the TEP Summary report may be posted.)
 - Posting for public comment.
 - Testing measures for reliability and validity, as well as ease and accuracy of collection.
 - Refining measures as needed.
- ◆ The Measure Information and Measure Justification forms including a summary of any changes proposed.
- ◆ A list of the TEP members, including any potential conflicts of interest disclosed by the members.
- ◆ Information about the measure contractor and subcontractors reevaluating this measure set.

Figure 13-1 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 site.
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.

Step 4: Collect information

Commenters will submit their comments via email or other Web-based tool 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 Public Comment Summary Report by summarizing and analyzing the public comments received. Preliminary recommendations may be stated here pending discussion with the TEP. This report should be submitted to the COR/GTL and the measure contractor's TEP within two weeks following the end of the public comment period. Upon approval by the COR/GTL, verbatim comments submitted will be made viewable to the public by posting on the CMS Web site. Work with the Measures Manager to post the report.

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.
- ◆ A summary of the comments for each measure and any recommended action.
- ◆ A 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.

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: Report on measure contractor's recommendations in response to comments

Finalize the Public Comment Summary report by documenting the TEP discussion and the recommended actions. Submit it to the COR/GTL within one week after the TEP meeting to review the comments. This step can also be done via email if only a few comments to the proposed measure(s) are received.

The report should include:

- ◆ The measure contractor's recommendations and actions taken in response to the comments received.

- ◆ Updated or revised measure specifications and justification with notations about changes made.

Step 8: Arrange for the 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 summary report and verbatim comments within three weeks (or as directed by the COR/GTL) after the public comment period closes. This posting will remain on the Web page 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.

Template for Call for Public Comment

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 each measure):*
- ◆ *<Measure A-Measure Name>*.
- ◆ *<Measure B-Measure Name>*.
- ◆ *<Measure C-Measure Name>* etc.

Template for Public Comment Summary Report

<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 measures)

Template for Public Comments Verbatim

Following this page is the file that can be used for documenting the verbatim public comments.



Public Comment Verbatim

Date Posted	Measure Set or Measure	Text of Comments	Name, Credentials, and Organization of Commenter	E-Mail Address	Type of Organization	Recommendations/Actions Taken

14. Measure Testing During Maintenance Phase

14.1 Introduction

Measure testing enables a measure contractor to monitor and 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 evaluation criteria on an ongoing basis. This information can be used in conjunction with expert judgment to evaluate a measure. Within the measure maintenance framework, measure testing is most likely to occur during a comprehensive reevaluation, but may also occur during an ad hoc review. Properly conducting measure testing and analysis is critical to continued 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 maintenance, the procedure for planning and testing under the direction of the 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. Similarly, the focus and extent of testing during maintenance differs based upon the evidence previously obtained prior to measure endorsement, and may also differ based upon changes to medicine, policy, and practice since measure implementation. Measure contractors should always select testing that is appropriate for the measure being monitored, and always provide empirical evidence for importance to measure and report, feasibility, scientific acceptability, and usability and use.

14.2 Deliverables

- ◆ Measure Testing Plan
- ◆ Measure Testing Summary report
- ◆ Updated Measure Information Form (MIF)
- ◆ Updated Measure Justification form
- ◆ Updated Measure Evaluation report

14.3 Extent of Measure Testing

Within the measure maintenance cycle, an endorsed measure must be reevaluated using the measure evaluation criteria described in the Measure Evaluation During Maintenance Phase section. The requirement and contribution of measure testing with respect to the reevaluation depends upon whether the measure has been implemented in a population (that is, if the measure is currently in use) and whether it has undergone a comprehensive reevaluation since the measure's initial endorsement. The extent of measure testing under these different circumstances is outlined in Table 14-1. This table

reflects NQF Task Force recommendations¹ acknowledging that evidence supporting importance, scientific acceptability, feasibility, and usability accumulates over time, and that evidence submitted during the measure endorsement process is not definitive. This view motivates the comprehensive reevaluation requirements shown in the first row of Table 14-1, which lists testing requirements for endorsed measures where an initial comprehensive reevaluation has not yet occurred. However, the NQF Task Force also recognizes that repeatedly demonstrating a measure’s scientific acceptability across multiple reevaluations every three years represents an undue burden. Consequently, as shown in the second row of Table 14-1, little or no measure testing may be required after the first comprehensive reevaluation has been completed.

Table 14-1 Extent of Measure Testing as Function of Prior Comprehensive Evaluation and Measure Use

	Measure in use	Measure not in use
First comprehensive reevaluation	Contractor should obtain data from the population where the measure was implemented, and analyze it to augment previous evaluation findings obtained when the measure was initially developed and endorsed.	Contractor should conduct expanded testing relative to the initial testing conducted during development (e.g., expand number of groups/patients included in testing compared to prior testing used to support the measure’s initial development and submission for endorsement).
Subsequent comprehensive reevaluation	If measure has not materially changed, CMS may require minimal testing and may use prior testing data for NQF maintenance if past testing results demonstrated a high rating for reliability and validity of the measure.	If measure has not materially changed, contractor may submit prior testing data when past results demonstrated adequate reliability and validity of the measure.

14.4 Testing For a Material Change

A material change is one where the modifications to a measure’s specifications affect the original measure’s conceptual framework or logic, the intended meaning of the measure, or the strength of the measure relative to the measure evaluation criteria. When a material change occurs, test the measure to ensure that it still meets evaluation criteria.

14.5 Alpha and Beta Testing

Testing may be required when revising a measure to ensure that a material change has not adversely affected performance with respect to the measure evaluation criteria. Testing during maintenance 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 planning for alpha or beta tests.

¹National Quality Forum, *Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties*, Washington, DC: NQF, 2011.

It is important to note that the testing does not need to address every aspect of the measure being revised or that is undergoing an initial comprehensive reevaluation. For a revised measure that has been materially changed, testing may not need to address the validity of data elements comprising the measure if these have already been thoroughly tested. Similarly, for an initial comprehensive reevaluation, NQF recommends that testing focus on measure performance and/or resulting accuracy of classification, rather than focus on data elements used in measure score calculations.

Alpha testing

Alpha tests are of limited scope. They usually occur during the outset of measure revisions before detailed specifications are finalized. The primary purpose of alpha testing during the maintenance of a measure is to refine or confirm the feasibility and utility of envisioned changes to technical specifications. The types of testing done in an alpha test vary widely, and often depend upon the measure’s data source or the specifications for the measure.

Beta testing

Beta testing generally occurs at the onset of the initial comprehensive evaluation or after the measure contractor has developed the detailed and precise updates to the technical specifications for a materially changed/revised measure. Beta tests serve as the primary means to assess scientific acceptability and usability of a measure after implementation and any updates to the measure’s technical specification. They can also be used to evaluate the measure’s existing risk adjustment/stratification methods, and help expand prior evaluations of the measure’s importance, usability, 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 14-2 compares key features of alpha and beta testing.

Table 14-2 Features of Alpha and Beta Testing

	Alpha Testing	Beta Testing
Timing	<ul style="list-style-type: none"> ◆ Usually carried out prior to the completion of revisions to technical specifications. ◆ May be carried out multiple times in quick succession. 	<ul style="list-style-type: none"> ◆ Occurs at the onset of the initial comprehensive evaluation or after the measure contractor has developed the detailed and precise updates to the technical specifications for a materially changed/revised measure.
Scale	<ul style="list-style-type: none"> ◆ Typically smaller scale. ◆ Only enough records to ensure data set contains all elements necessary to evaluate the specification revisions. ◆ Only enough records to identify common occurrences or variation in the data. 	<ul style="list-style-type: none"> ◆ 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).

	Alpha Testing	Beta Testing
Sampling	<ul style="list-style-type: none"> ◆ Convenience sampling. 	<ul style="list-style-type: none"> ◆ 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	<ul style="list-style-type: none"> ◆ Permits the early detection of problems in the revised technical specifications (e.g., identification of additional inclusion and exclusion criteria). 	<ul style="list-style-type: none"> ◆ Used to assess or revise the complexity of computations required to calculate the measure.

	Alpha Testing	Beta Testing
Importance	<ul style="list-style-type: none"> ◆ 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.). ◆ Additional evaluation of potential gaps in current measures that the new measure is attempting to address. ◆ Provides support for further development of the measure. 	<ul style="list-style-type: none"> ◆ Allows for enhanced evaluation of a measure's 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.	<ul style="list-style-type: none"> ◆ Assesses measure reliability and validity. ◆ Assesses the impact of a measure revision upon prior findings. ◆ Report results of analysis of exclusions (if any used). ◆ Test results of risk adjustment model, quantifying relationships between and among factors.
Feasibility	<ul style="list-style-type: none"> ◆ Identifies barriers based on implementation of the measure. ◆ 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. 	<ul style="list-style-type: none"> ◆ 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		<ul style="list-style-type: none"> ◆ Describe data and results if usefulness demonstrated during testing

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.
- ◆ Be randomly selected, when possible.

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.

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>.

14.6 Procedure

When reassessing a measure for CMS, a measure contractor 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 contractors 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 maintenance contract.

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 should include sufficient information to help the COTR/GTL understand how the sampling and planned analyses ensure that the measure or set of measures meet scientific acceptability, usability, and/or feasibility criteria required for approval by CMS and continued endorsement by NQF.

The testing plan may contain some or all of the following.

- ◆ Measure name(s)
- ◆ Type of testing (alpha or beta)
- ◆ Study objective(s)
- ◆ Timeline for 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, or 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 CMS information collection policy.
- ◆ 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

Analyze the testing results following the analysis plans specified in the work plan. Discuss any modifications or deviations from the intended analysis with the COR/GTL as needed.

Step 5: Refine the measure

For a revised measure that has been materially changed, additional modification may be required based on the testing results. For a measure undergoing a comprehensive reassessment, changes may be required if testing detects problems or suggests that scientific acceptability criteria have not been met (e.g., changes in the definition of the population may be required based upon unforeseen issues following full implementation of the measure in the target population).

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 changes in importance (e.g., changes in variability across comparison groups), reliability, validity, and exclusion results; changes to risk adjustment or stratification methods; 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 sub criteria. It is important to evaluate the measure in an as objective manner as possible in order to anticipate any issues when the measure is reevaluated for endorsement by NQF. 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 four NQF measure evaluation criteria (i.e., Importance to Measure and Report, Scientific Acceptability of Measure Properties, Usability, and Feasibility), the goal of these summaries is to document sufficient evidence to achieve approval by CMS and endorsement by NQF.

Not all revisions will require extensive reassessment of all testing criteria. Not all previously endorsed measures will meet each criterion equally. This is often a matter of judgment and expertise. Given the complexity of assessment, measure contractors need to work with experienced statisticians and methodologists to provide expert judgment when reporting measure reliability and validity. Contractors should 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

Maintenance testing information will be for the NQF 3-Year Maintenance Review and will be included in the Measure Maintenance form.

14.7 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 importance and feasibility of potential changes in a measure's specifications.

- ◆ The findings from one or more beta tests are often used to demonstrate scientific acceptability and usability of measure specification modifications or performance following implementation, 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, additional testing results can be used to demonstrate that a measure continues to assess an area with substantial opportunities for improvement. Testing following implementation of a measure 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 regarding the intended area of measurement. These qualities determine whether the measure can be used to draw reasonable conclusions about care in a given domain.

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 conditions. 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 contractor will work with 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 monitoring 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

For a comprehensive reevaluation, NQF recommends that testing focus upon measure performance, rather than data elements.² However, for materially changed measures, testing of new data elements may still be warranted, especially if the new/revised elements contribute heavily to the computed measure score.

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—Inter-rater 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—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

²National Quality Forum, *Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties*, Washington, DC: NQF, 2011.

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—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—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.³ 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 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 and maintenance, 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).

³Cronbach’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. *Psy- chometrika* 16:297–334.

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.⁴

- ◆ Construct validity—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—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—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 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—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—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—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

⁴Messick, 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.

participants systematically rate their agreement, and formal aggregating and consensus failure processes are followed.⁵ 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.
- ◆ 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.

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 or revised elements that contribute most to the computed measure score.

Prior evidence of reliability and validity for measure elements

Unlike a measure score that requires reevaluation, testing is not always necessary for data elements comprising the score when prior evidence of reliability or validity exists; and this evidence can often be used in place of testing data elements during the reevaluation. Prior evidence can include published or unpublished testing. NQF⁶ provides the following guidance:

- ◆ Validity—Prior evidence of data element validity may be used provided it uses the same data elements and data type as the measure being monitored, and was obtained using a representative sample of sufficient size. Data elements that represent an existing standardized

⁵Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties. (2011): The National Quality Forum—Task Force on Measure Testing.

⁶Guidance 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.

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 data element reliability may be used provided it uses the same data elements and same data type, and was obtained using a representative sample of sufficient size.

Testing of exclusions/exceptions

Exclusions/exceptions used in a measure’s specifications should be reexamined during the comprehensive reexamination. Changes may have occurred following measure implementation, or additional data now available may detect issues that were not apparent during testing conducted to support initial measure endorsement. 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.
- ◆ 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

Measure testing may be used to evaluate the continuing adequacy of the risk adjustment strategy when the measure is an outcome measure. Risk adjustment models should be recalibrated regularly to ensure the estimated coefficients remain appropriate for the measure over time. Risk adjustment may not be necessary when the measure being revised 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.

Usability

Formal usability testing is often not required for a comprehensive reevaluation or a revised measure, 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

Widely-used measures generally do not require reexamination of feasibility unless public comment or other feedback suggests problems. Similarly, changes to a measure may not substantially affect a measure’s feasibility, and this aspect of testing may be omitted if the argument for unchanged feasibility can be justified. If testing is warranted, testing data may be used to assess measure feasibility to determine the extent to which the required data are available and 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.

For measures in use or that have been materially changed, feasibility assessments can address the following:

- ◆ Changes to the availability of data (e.g., evidence that required data, including any exclusion criteria, is routinely generated and used in care delivery).
- ◆ Changes to the extent of missing data, measure susceptibility to inaccuracies, and the ability to audit data to detect problems.
- ◆ An estimate of changes to cost or burden of data collection and analysis.
- ◆ Any barriers encountered in implementing performance measure specifications, data abstraction, measure calculation, or performance reporting.
- ◆ Changes to 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 compromise confidentiality.
- ◆ Identification of unintended consequences.

14.8 Special Considerations



Testing measures specified for use in EHRs (eMeasures)

A health quality measure (or clinical quality measure) encoded in the Health Quality Measures Format (HQMF) is referred to as an “eMeasure.” It is a standard for representing a health quality measure as an electronic document. The HQMF provides for quality measure consistency and unambiguous interpretation through standardization of a measure’s structure, metadata, definitions, and logic. These requirements are intended to promote measures that are reliable and valid when extracted across diverse electronic health record systems. However, these eMeasures will still require adequate testing to ensure reliability, validity, and also feasibility of accurately extracting the measure from the diverse EHRs in which the measure has been (or will be) implemented.

Many different EHR systems are in use today (particularly in the ambulatory care setting), so it is unlikely that initial measure testing would include all installations and vendor systems. Once an eMeasure has been implemented in diverse EHR systems, this provides opportunities for ongoing evaluation and testing of the eMeasure during comprehensive reevaluation and NQF’s measure maintenance process. This is particularly true given that (1) some EHR systems may not have been available at the time of initial testing (prior to endorsement), and (2) eMeasure testing can only provide a snapshot of EHR capabilities, and (3) advancements to vendor systems occur on a continual basis. Testing after implementation also offers the opportunity to assess the impact of the diverse provider practice patterns on the measure.

For measures not in use, reevaluation offers an opportunity to expand on previously conducted testing, perhaps expanding the number of settings or EHR systems used relative to the originally conducted measure testing. When recruiting sites and vendors to test an eMeasure, measure contractors should consult with their COR/GTL regarding the Paperwork Reduction Act of 1995, which requires Office of Management and Budget (OMB) approval before requesting most types of information from the public.

eMeasures with material changes to specifications may also require additional testing. Examples of material changes would include those affecting the original measure’s concept, logic, intended meaning, or strength of the measure relative to the measure evaluation criteria. Some strategies for testing during the comprehensive reevaluation process are discussed below.

Validity testing

Ideally, certified EHRs will record clinical information in discrete computer-readable fields, which potentially reduces errors when extracting data elements for measure calculation and reporting. However, even under these circumstances, measures may need to be reevaluated during measure testing. For example, a measure may be impacted by the complexity of the specifications, such that it is susceptible to differences in usage of the data fields by different users or it may incorporate elements that are frequently entered into the wrong EHR fields. A measure may also be impacted simply due to small changes in the clinical vocabulary code sets used. Given the difficulty of ensuring standardized

data extraction across diverse EHRs and EHR users, different options should be considered when testing an eMeasure during ongoing maintenance periods. NQF recommendations for retooled eMeasures (and time-limited measures)- previously tested on the original data source- includes testing of reliability and validity by the next endorsement maintenance review using one of the methods below⁷. These same methods should be considered for de novo measures undergoing maintenance.

Comparison to abstracted records—eMeasure contractors should compare measure scores and measure elements across multiple data sources, particularly 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 in assessing the measure’s susceptibility to differences in EHR applications or data entry practices. Standard assessment of the agreement between the two methods using a reference strategy/criterion validity approach is often sufficient to demonstrate comparability. For measures that have been implemented prior to the comprehensive reevaluation, random sampling from the population in which the measure has been implemented is acceptable, but restricted sampling strategies that ensure representation from most/all EHR systems may also be beneficial. For measures that are not in use, expanded testing available through recruitment of volunteer test sites may be sufficient.

Comparison to simulated QDM data set—For measures not yet in use or newly implemented, measure contractors may have difficulty obtaining a sample that adequately represents the different practice patterns and certified EHR systems that will be in use when the measure is fully implemented. To address this issue, validity testing may 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

NQF has developed assessment criteria for eMeasure validity and reliability that encompasses examination of measure properties using both the authoritative source (e.g., manually reviewed charts) and a simulated data set approach described in the previous section.⁸ Although the 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.⁹ For *measure data elements*, validity is demonstrated if either (a)

⁷National Quality Forum eMeasure Memo from Heidi Bossley to CSAC on March 6, 2012: *Review of Comments on NQF Requirements for eMeasure Review and Testing*. Available at http://www.qualityforum.org/About_NQF/CSAC/Meetings/2012_CSAC_Meetings.aspx, Accessed: May 9, 2012.

⁸National Quality Forum. *Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties*. NQF Task Force on Measure Testing. 2011:50.

⁹National Quality Forum. *Draft Requirements for eMeasure Review and Testing*. NQF November 2011. Available at http://www.qualityforum.org/.../Draft_Requirements_for_eMeasure_Review_and_Testing.aspx. Accessed March 26, 2012.

adequate agreement is observed between electronically and manually extracted data elements abstracted from the entire EHR, or (b) 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. Consequently, eMeasures undergoing a comprehensive reevaluation do not necessarily require testing of the measure data elements, if prior evidence demonstrates sufficient validity. However, when an eMeasure has been widely implemented and element-level data is available from a large, diverse number of providers, examining differences across groups may help determine if provider practice patterns or specific EHR systems result in dramatic differences in availability of data element level information specified in the measure. This examination may also help inform recommendations for changing the measure specifications (e.g., dramatic differences in the storage location of specific elements based on the particular EHR software used may suggest areas that should be reviewed if measure specifications are changed).

Feasibility of eMeasures

While usability of an eMeasure generally is demonstrated in a manner equivalent to a non-eMeasure, 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. For measures that have not yet been implemented at the time of the comprehensive reevaluation, the prior review may be sufficient, or it may be augmented by evaluation across a larger set of EHR systems. For measures that have already been implemented, feedback from eMeasure implementers may provide additional information suitable for the reevaluation.

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 contractor 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

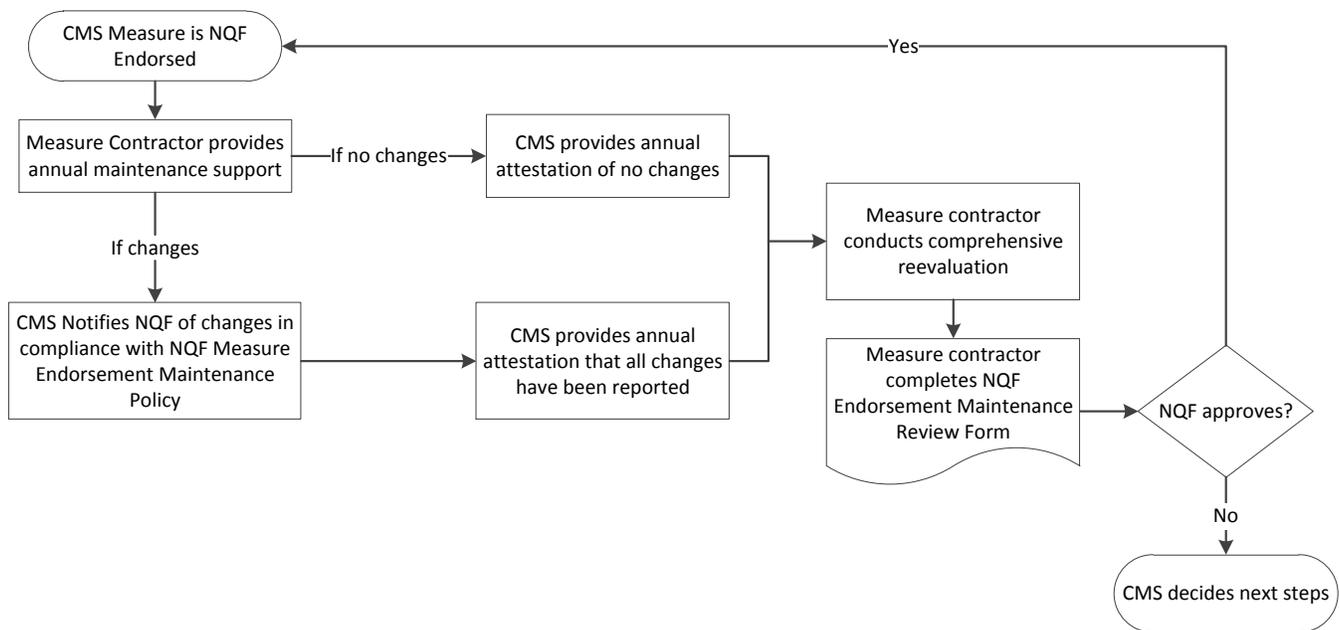
Measure testing may also demonstrate that the measure can be consistently implemented across organizations by quantifying comparable variation for individual components, and demonstrate that the measure can be decomposed into its components at the group/organization level to facilitate transparency and understanding by the intended measure audience.

15. National Quality Forum Endorsement Maintenance¹

15.1 Introduction

Once the National Quality Forum (NQF) has endorsed a measure, the measure contractor supports ongoing maintenance of the endorsement of the measure if it is part of the scope of work for that contractor. This section describes NQF’s standardized measure endorsement maintenance processes. Figure 15-1 illustrates the key steps in this process. Measure endorsement maintenance processes are subject to change as deemed appropriate by NQF. The measure contractor is responsible for being familiar with NQF’s current measure endorsement maintenance processes described on NQF’s Web site at: http://www.qualityforum.org/Measuring_Performance/Maintenance_of_NQF-Endorsed®_Performance_Measures.aspx

Figure 15-1 NQF Maintenance Process



15.2 Procedure

NQF has developed a set of endorsement maintenance processes and reviews to ensure that endorsed measures continue to meet the endorsement criteria. NQF endorsement maintenance reviews are separate from the Measure Management System maintenance reviews. The relationship of the Measures Management System maintenance reviews and NQF endorsement reviews are described in

¹The 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.

this section. Figure 15-2 below also shows the relationship of measure maintenance reviews and NQF endorsement maintenance reviews.

NQF conducts three types of measure endorsement maintenance reviews.

- ◆ Annual update—these are usually done annually beginning with year 1 (year 0 is considered the year when endorsement was granted).
- ◆ 3-year endorsement maintenance review—this is done every 3 years beginning with year 3 (year 0 is considered the year when endorsement was granted).
- ◆ Ad hoc endorsement review—this is done only when requested by any interested party or as deemed necessary by NQF.

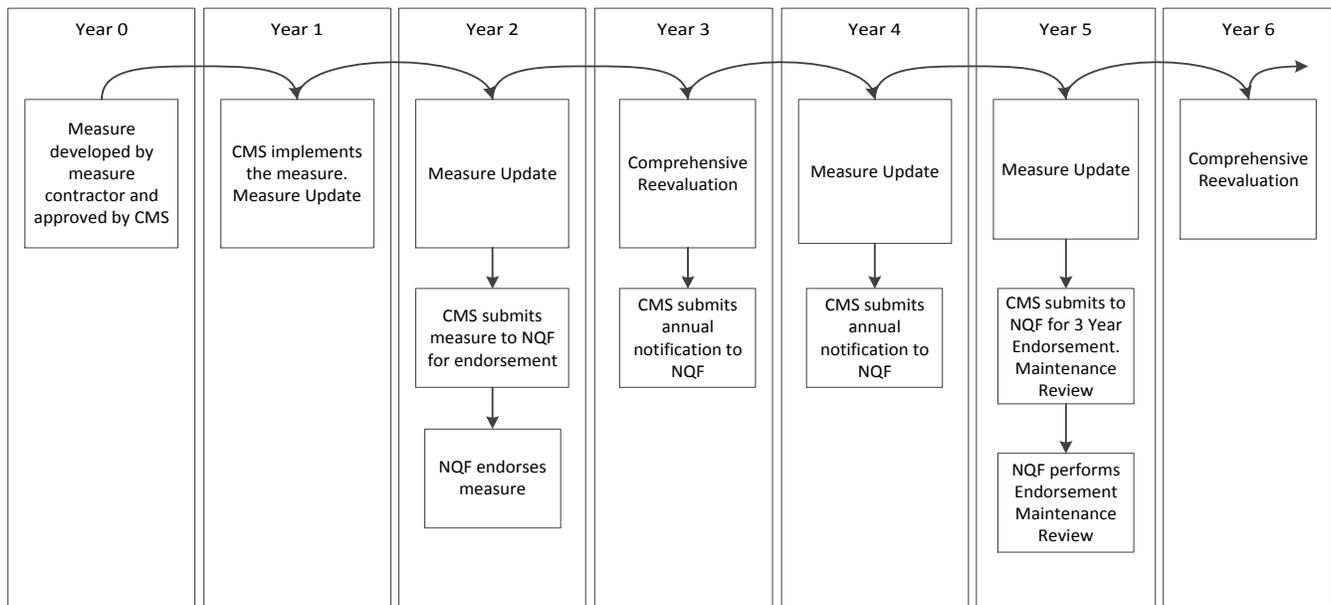


At this time, the inclusion of eMeasures for all measures going through NQF maintenance is optional.² Should a developer choose to submit an eMeasure, it should be understood that they are required to undergo maintenance every three years as well. If the measure developer is contractually required to provide measure maintenance support, the contractor will support the Centers for Medicare & Medicaid Services (CMS) during these endorsement maintenance reviews. Measure contractors should follow NQF's online measure submission processes for measure endorsement maintenance. The contractor can request a free user account to file the measure submission forms and gain access to the measure dashboard at the following location:

<http://imis.qualityforum.org/Core/CreateAccount.aspx>

²National Quality Forum: *Health Information Technology Advisory Committee (HITAC) Meeting Summary* May 22, 2012.

Figure 15-2 CMS Measure Reevaluation and NQF Measure Endorsement Maintenance



NQF annual status update report

After a measure is NQF-endorsed, CMS, as the measure steward, is required to submit a status report of the measure specifications to NQF annually. The measure contractor is responsible for preparing this annual update report and obtaining the Contracting Officer Representative/Government Task Leader (COR/GTL) review and approval before submitting it to NQF. The measure contractor will update their measures following the CMS Measures Management System process for Measure Update process described. Some measure maintenance contracts may require measure updates more than once a year. In those cases, measure contractors should notify NQF of the changes as often as appropriate. Otherwise, only one annual update is required.

NQF provides a standardized online submission template for annual measure maintenance. The annual status update report focuses on measure specifications. This report either affirms that the detailed measure specifications of the endorsed measure have not changed; or if changes have been made, the report documents the details and underlying reasons for the changes. In some instances, upon consideration, NQF staff may decide that the changes made necessitate an ad hoc review of the measure.

NQF allows measure developers to select the quarter in which they wish to submit their annual updates. The deadline for each measure update should be confirmed with NQF annually along with contact information. The updates also appear on measure contractor’s NQF dashboard. It is the responsibility of the measure contractor to track when updates are due and visit their NQF dashboards periodically. Note that NQF sends email notifications only for 3-year comprehensive maintenance reviews and Time Limited Endorsement measures but not for annual updates. CMS and the Measures Manager are notified only when the measure contractor responsible for measure maintenance fails to

meet the deadline and have outstanding updates. It is the responsibility of the maintenance contractor to ensure the updates are submitted in a timely manner.

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.

If the measure contractor is not able to get the information needed in a timely manner and anticipates being late submitting the form, the contractor should notify the COR/GTL as soon as possible. The COR/GTL will report the situation to NQF and may seek extension of the due date. The measure contractor and COR/GTL may seek guidance from Measures Manager during any stage of this process.

NQF endorsement maintenance—3-year maintenance review

NQF requires measures that have been previously endorsed to be re-evaluated for continued endorsement every three years. The endorsed measures are reevaluated against NQF's Measure Evaluation Criteria and are reviewed alongside not yet endorsed measures. This concurrent comparison of newly submitted and previously endorsed measures fosters harmonization and helps ensure that NQF is endorsing the best available measures. 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.

Measure contractors who are responsible for the maintenance of CMS-developed measures will conduct comprehensive reevaluations as described in this volume. The CMS Measures Management System requires measures to be reevaluated every three years. The information gathered for these reevaluations should be used to complete NQF maintenance submissions. The timing of the comprehensive reevaluations can be coordinated with the anticipated timing of NQF endorsement maintenance reviews. The Comprehensive Reevaluation ideally, should precede NQF scheduled review, so that CMS, along with the measure contractors, can determine the outcome of the reevaluation and address any harmonization issues identified. The time required for testing of any significant changes made to the measures should be factored into the timing of the Comprehensive Reevaluation.

NQF will notify CMS of the upcoming expiration of a measure's NQF endorsement six months prior to that date. The notification will also appear on the measure contractor's dashboard. The Measure Manager or the COR/GTL will confirm with the appropriate measure contractor that they received NQF notice. Even though NQF sends reminders and email notifications about the maintenance review due date, measure contractors are responsible to be aware of NQF endorsement expiration dates and seek advice from their COR/GTL or NQF if they have not received notification of an endorsement maintenance review.

Measure contractors should follow the standardized online submission form template for the three-year endorsement maintenance review. This form will be prepopulated with measure information from the most recent submission. The three-year endorsement maintenance review form should document review of current evidence and guidelines, and provide information on how the measure still meets the criteria for NQF endorsement. This information should be readily available from the contractor's own Comprehensive Reevaluation of the measure and subsequent environmental surveillance. Measure contractors will also provide information regarding any modifications to the measure specifications, data supporting use of the measure, testing results, and other relevant information.

The measure contractor should be aware of NQF's testing requirements for first maintenance review and subsequent reviews. These requirements are discussed in the Volume 2 Measure Testing During Maintenance Phase section.

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 measure contractor is responsible for updating the information in the form and for obtaining COR/GTL review and approval before submission 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.

It is the measure contractor and COR/GTL's responsibility to make sure that the form is submitted before the expiration due date. If a measure contractor is not able to get the information (testing data etc.) needed to complete the form in a timely manner and anticipates being late, the contractor should inform the COR/GTL as soon as possible. The COR/GTL will report the situation to NQF and may seek extension of the due date. The measure contractor and COR/GTL may seek guidance from the Measures Manager during any stage of this process.

Measure contractors are advised to attend the Steering Committee and Consensus Standards Approval Committee (CSAC) meetings during endorsement maintenance review as they were during the initial review. Contractors should be prepared to answer any questions about the measure that the committee, CSAC Board, or NQF staff members might have.

Measure maintenance—ad hoc review

NQF may conduct an ad hoc endorsement review on an endorsed measure at any time if any of the following occur:

- ◆ Evidence supporting the measure has changed.
- ◆ Implementation of the measure results in unintended consequences.

- ◆ Material changes have been made to the measure which may affect the original measure's concept or logic.

Ad hoc endorsement reviews may be requested at any time by any party. Adequate evidence to justify the review and under which criterion the review is requested, must be submitted in writing when seeking an ad hoc maintenance review. NQF reviews the request and initiates an ad hoc review as long as there is adequate evidence to justify one. The timing of the ad hoc review will be determined by the presence of any accompanying safety concerns associated with the changes to the endorsed measure.

If NQF has received a request for an ad hoc maintenance review, NQF will notify the steward whether or not NQF has determined that there is sufficient evidence to conduct the ad hoc review. If an NQF ad hoc review is requested for a measure supported by the measure contractor, the contractor is responsible for assisting CMS in response to the request from NQF. NQF currently does not have a form for the ad hoc review. The measure contractor and CMS should meet with NQF to discuss the request and clarify the types of information that should be submitted and the timeline for the ad hoc maintenance review.

Below are some scenarios and the measure contractors' role.

- ◆ NQF has received a request for review and has declined to conduct an ad hoc review.
 - CMS may request that the measure contractor conduct its own ad hoc review to further explore the concern raised. As described in the Ad Hoc Review section, the results of the review may be to retain the measure as is, revise the specifications, retire the measure or suspend the measure.
 - If CMS does not request the contractor to conduct an ad hoc review, the measure contractor should continue to monitor the measure and literature for additional evidence and consider the information during the next comprehensive review.
- ◆ NQF has received a request for a review and has determined that there is sufficient evidence to initiate the ad hoc maintenance review.
 - CMS will request that the measure contractor conduct its own ad hoc review to further explore the concern raised. As described in the Ad Hoc review section, the results of the review may be to retain the measure as is, revise the specifications, retire the measure or suspend the measure.
 - If the outcome of the measure contractor's ad hoc review is to retire the measure, CMS will notify NQF and NQF ad hoc maintenance review will not be necessary.
 - If the outcome of the measure contractor's ad hoc review is to revise the measure and include supporting evidence, the revised specifications and justification will be submitted to NQF and reviewed by NQF.
 - CMS requests an ad hoc review. CMS may request an ad hoc review of its measures if it makes a significant change to a measure and wants to ensure the continued endorsement of the measure. Examples of changes considered significant include
 - A significant revision of a measure may have been made as a result of an ad hoc review initiated by CMS or the measure contractor.

- The measure specifications have been expanded for use in additional populations, and there is a need to expedite NQF's review due to legislative or program requirements.

The ad hoc review process follows a shortened version of the consensus development process and includes a technical expert call for nominations, expert panel review, a public and member comment period of no less than 10 days, CSAC review, NQF Board of Directors ratification, and an appeals period. This time period is subject to change depending on the nature of the submitted measure.

Measure contractors are advised to attend the CSAC meetings during the ad hoc endorsement review as they were during the initial review. Contractors should be prepared to answer any questions about the measure that the committee, the board, or NQF staff members might have.

If a measure remains endorsed after ad hoc review, it is still subject to its original maintenance cycle.

15.3 NQF 2-Stage Consensus Development Process (CDP)

Currently NQF is running a pilot project to assess the feasibility of its newly designed 2-stage CDP. It is anticipated that first measure reviews using the new process would begin sometime in 2013. Information regarding this process is described below. This new process does not impact the endorsement maintenance process, and is provided in this section for informational purposes only.

Stage 1

In this stage for measure maintenance, the measure developers would submit their previously endorsed measures, as concepts, to NQF. Developers can ask NQF staff to review forms to ensure for accuracy and completion 30 days prior to submission deadline. The concept of the endorsed measures would be evaluated by the Steering Committee against NQF's Importance to Measure and Report criterion. Per NQF, Stage 1 review will also help identify harmonization and competing measure issues. Thereafter, the measures that pass the importance criterion will go to CSAC and the Board for approval. Once NQF has approved the importance criterion of the measure concepts and any harmonization issues have been addressed, they then move into Stage 2 of the CDP.

Another important thing for measure developers to note is that under the new process, the plan is for the 19 steering committees to review the measure concepts more frequently. Instead of waiting for the current three-year cycle, measure developers would have opportunities to submit their measures twice a year. These steering committees plan a preset schedule throughout the year to evaluate measures and 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.

15.4 Conclusion

Maintaining NQF endorsement of its quality measures is of high importance to CMS. It is the responsibility of measure contractors and COR/GTLs to work together to help maintain the endorsement and stay abreast of the most current NQF endorsement policies and procedures. Measure contractors are also responsible for timely, accurate, and complete submission of relevant forms.



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.

¹Krumholtz, H., et al. Standard for measures: efficiency in health care. *Journal of the American College of Cardiology*. 2008;52 (number 18): p. 1522.

²Leatherman, S., Berwick, D., et al. The business case for quality: case studies and an analysis. *Health Affairs*. 2003; 22, (number 2): p. 18.

³National Committee for Quality Assurance. *IOM’s Envisioning the National Health Care Quality Report*, 2001. Appendix E.

⁴National Research Council. *IOM’s Envisioning the National Health Care Quality Report*, 2001. Appendix E.

⁵National Quality Forum, *Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety: Composite Measures, Draft Report*. July 2009, page 2.

⁶The 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.

⁷Truman 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.

⁸National Quality Forum. *Quality Data Model Description*. Available at: http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx. Accessed July 30, 2012.

⁹Merriam-Webster dictionary; Available at: <http://www.merriam-webster.com/dictionary/de%20novo>

¹⁰Truman 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.

¹¹Leatherman, et al. p. 19.

¹²Agency for Healthcare Research and Quality. *Health Care Efficiency Measures*. Available at: <http://www.ahrq.gov/qual/efficiency/>. Accessed March 30, 2011.

¹³National Quality Forum (NQF). *Measurement Framework: Evaluating Efficiency Across Patient-Focused Episodes of Care*. Washington, DC: NQF; 2009.

¹⁴American 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.

¹⁵McGlynn EA. Selecting Measures of Quality and System Performance. *Medical Care*. 2003;41(suppl) 39-47.

¹⁶Available at: <http://www.qualityforum.org/MAT/>. Accessed August 1, 2012

¹⁷National Quality Measures Clearinghouse Glossary. Available at: <http://www.qualitymeasures.ahrq.gov/about/glossary.aspx>. Accessed June 13, 2011.

¹⁸Overview of Risk Adjustment and Outcome, Measures for Home Health Agency OBQI Reports: Highlights of Current Approaches

¹⁹National 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.

²⁰National 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

²¹Merriam-Webster’s Online Dictionary, Public Domain. Available at: <http://www.merriam-webster.com/dictionary/public+domain>. Accessed June 5, 2012.



²²National Quality Forum. *Quality Data Model Description*. Available at: http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx. Accessed August 1, 2012.

²³Ibid.

²⁴Ibid.

²⁵Leatherman, et al. p. 19.

²⁶National Quality Forum. *Quality Data Model Description*. Available at: http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx. Accessed August 1, 2012.