Report from the Intended Use Advisory Panel: Preliminary Recommendations to Enhance the Consensus Development Process (CDP)

A CONSENSUS REPORT

August 31, 2015

Contents

Preliminary Recommendations to Enhance the Consensus Development Process (CDP)

DRAFT REPORT

Executive Summary

The National Quality Forum (NQF) strives to continuously improve its processes to reflect the changing needs of patients and the American healthcare system in which they get their care. Since its inception, one of the central principles of NQF's endorsement process has been that endorsed measures, having met rigorous criteria and approval from diverse stakeholders, should be suitable for use in both quality improvement and accountability (i.e., payment or public reporting) applications. However, over the past several years, a number of stakeholder groups have questioned whether NQF's endorsement process should consider the specific intended or actual use(s) of a measure.

This issue was considered by NQF's Consensus Task Force, which reviewed the endorsement process and recommended several specific enhancements to that process. Acting on the recommendations of the Consensus Taskforce, NQF convened an Intended Use Advisory Panel in December 2014 to examine (1) whether and how the specific use of measures should be incorporated into the endorsement process; and (2) whether endorsement decisions should change from a simple yes/no to an incremental scale that would allow NQF to apply the same criteria to all measures, but to grade endorsed measures differently based on how well the measures met the criteria and, potentially, how well the measures performed when implemented. The Advisory Panel is seeking broad input from the public on its recommendations to enhance the consensus process. Reflecting this input, a revised version of this report will be sent to the Consensus Standards Approval Committee (CSAC) and NQF Board for their consideration.

A number of themes emerged from the Intended Use Advisory Panel's discussions. All members of the Panel agreed that measures that are endorsed by NQF should produce reliable and valid information, regardless of the intended use of the measure. The Advisory Panel also agreed that providing more transparency to the public about the degree of a measure's reliability and validity would be very positive. Consequently, they recommended that endorsed measures be graded.

There was, however, limited consensus among panel members on whether the various accountability applications could be put into a hierarchy based on the risk of misclassification and the subsequent financial impact on providers. Consequently, the panel did not agree that the various grades assigned to measures should correspond to a single accountability application, e.g., some on the panel argued that AAA (the top rating with respect to validity and reliability) could apply to both pay for performance and public reporting applications. In sum, the Advisory Panel's recommendation is to provide grading for measures, thereby increasing transparency, but to not associate a particular grade with a particular application. The Advisory Panel encourages the Measure Application Partnership (MAP) to consider whether the grades could be used in the selection of measures for various programs. Along these lines,

the Panel recommended further examination of key measure and program attributes and their interaction to help inform both MAP and program implementers.

The Advisory Panel also noted that field experience could be a useful addition in the evaluation of measures used in accountability applications and that incentives should be put in place to encourage developers to evolve measures to higher grades. ,

While the Advisory Panel did not agree that it was appropriate to rank accountability applications, there was general agreement that there is a qualitative difference between measures used for quality improvement (QI) only and those used in other applications. Panel members suggested that different criteria for QI-only measures may help to create an environment where healthcare organizations can more easily use and share measures, while still retaining a role for NQF endorsement of QI measures to help avoid re-invention by end-users (including providers, plans and States).

To recap, the Intended Use Advisory Panel made the following recommendations:

1. Move forward with grading measures based on the NQF endorsement criteria
2. Incorporate “field experience” in the measure endorsement process
3. Consider the rapidly evolving uses of measures and incentivize grade advancement through further use and testing
4. Consider different endorsement standards for QI measures
5. Encourage the Measures Applications Partnership (MAP) to consider whether measure grades can be used in the selection of measures for programs
6. Pursue future work to consider the interaction between program attributes and individual measure attributes

The Advisory Panel seeks broad public comment on these key themes and the recommendations that emerged from its deliberations.

Introduction

The National Quality Forum (NQF) strives to continuously improve its processes to reflect the changing needs of the American healthcare system. A central component to NQF's role has been to endorse national performance measures that reflect rigorous criteria and input from stakeholders across the healthcare industry to support quality improvement and accountability applications. The value of NQF endorsement of measures stems from a rigorous and transparent evaluation process that is conducted by multi-stakeholder committees comprised of clinicians and other experts from the full range of healthcare providers, employers, health plans, public agencies, community coalitions, and patients—many of whom use measures on a daily basis to ensure better care. NQF-endorsed measures undergo routine "maintenance" (i.e., re-evaluation) to ensure that they are still the best-available measures and reflect the current science.

Over the past several years, several stakeholder groups have questioned whether endorsement should consider the specific intended or actual use(s) of a measure in its evaluation (e.g., in particular federal quality programs), and more broadly, whether the specific use of a measure should be included in the NQF measure endorsement criteria. This would move beyond the current use and usability measure endorsement sub-criterion which evaluates the extent to which potential 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, efficient healthcare for individuals or populations. This effort by the NQF Intended Use Advisory Panel seeks to consider the merit of and the various approaches to considering a measure's specific intended or actual use(s) as part of the measure endorsement process.

Background

Consensus Task Force

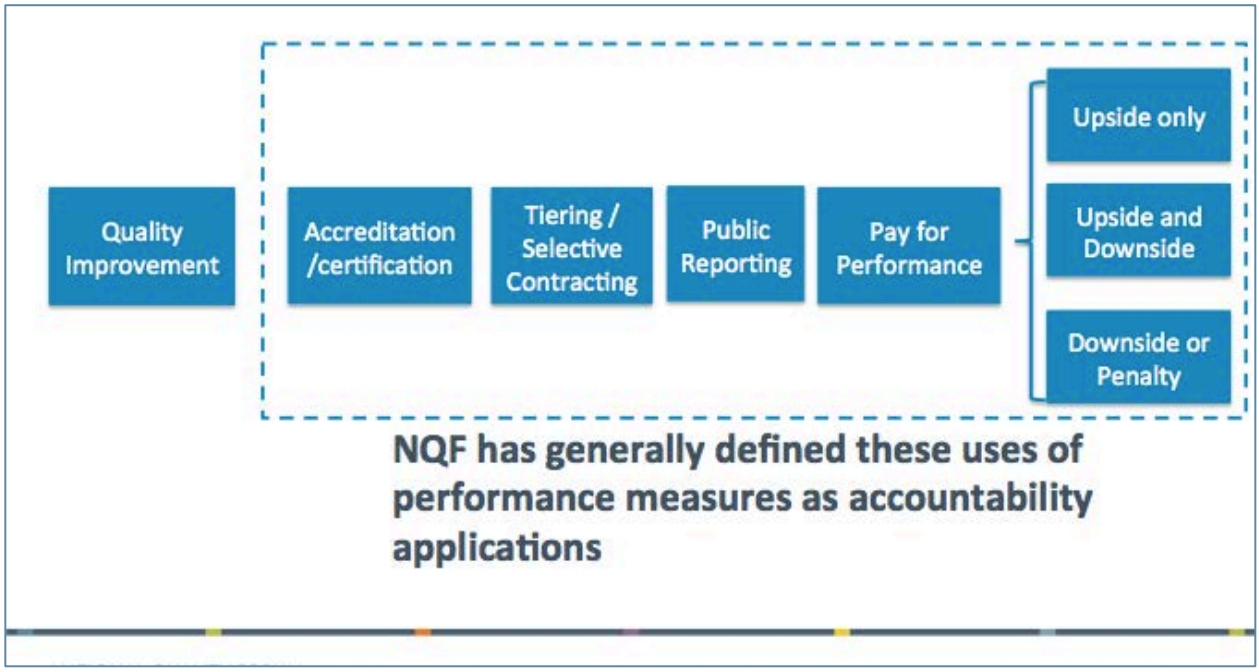
In 2012, NQF's Board of Directors (BoD) empaneled a Consensus Task Force (CTF) that included members of the NQF BoD, the Consensus Standards Approval Committee (CSAC), and individuals from the NQF membership. During their two-year tenure, the CTF reviewed the NQF endorsement process and recommended several specific enhancements to the process. As part of its final set of recommendations, the CTF, in the Fall of 2014, advised the NQF BoD to convene an Advisory Panel to consider transitioning from a binary endorsement decision (endorsed/not endorsed) to a more nuanced recommendation of endorsement. The CTF recommended that this Advisory Panel consider two potential new directions:

1. **Endorsement of measures for a specific intended or actual use(s)**: Currently, NQF endorses measures for quality improvement and accountability applications (see figure 1). Accountability applications have been generally defined as the use of performance measures not solely for the purpose of internal quality improvement. One potential direction recommended by the CTF is to consider endorsement for specific purposes (e.g., internal quality improvement, public reporting, payment), with consideration that measures may not be suitable for all potential uses. Endorsement for intended use would potentially allow NQF to hold measures used for

different purposes to different standards and would recognize that different stakeholders may have different priorities.

2. **Levels of endorsement by measure grading** (independent of use): Levels of endorsement would allow NQF to apply the same criteria to all measures, but to grade endorsed measures differently. Endorsement would change from a simple yes/no to an incremental scale; measures could move up or down the scale with periodic evaluations to maintain endorsement (perhaps even in less than the current three-year cycle) based on additional testing or experience. Measures could be graded based on how well they meet NQF endorsement criteria, along with experience.

Figure 1: The Various Uses of NQF-endorsed Performance Measures



The Intended Use Advisory Panel

Acting on the recommendations of the CTF, in December 2014 the NQF BoD approved the convening of an Intended Use Advisory Panel. Members of this Panel were selected to represent expertise in all aspects of quality measurement, including measure developers, program implementers, providers, and others who are measured by and use NQF-endorsed performance measures.

The Intended Use Advisory Panel was charged with:

- Discussing several critical topic areas, including identifying various use cases for NQF-endorsed measures, distinguishing among the use cases, and identifying the need, if any, for different measure attributes, depending on the specific intended or actual measure use(s);
- Examining if the NQF measure endorsement criteria requires updating;
- Proposing a path forward on whether, and if so, how, to incorporate the specific use of measures in the endorsement process.

Accordingly, this Advisory Panel did not design new processes nor develop new policies for NQF. Instead, their contributions should be viewed as foundational and strategic recommendations that will initiate and facilitate continued dialogue with NQF members and other stakeholders on this topic via various public commenting opportunities. The Advisory Panel prior to finalizing its recommendations and presenting them to the CSAC and NQF BoD will consider comments received.

The remainder of this report summarizes the key themes that emerged from the deliberations of this Advisory Panel and its initial recommendations.

Key Themes

The Intended Use Advisory Panel met via three web-meetings, as follows:

- June 8 & 10, 2015: Oriented to panel charge and key considerations outlined by the NQF BoD
- July 10, 2015: Considered the various uses for NQF-endorsed performance measures
- July 29, 2015: Considered how NQF endorsement criteria might vary based on the various use cases

The key themes that emerged from the Advisory Panel deliberations are summarized below.

Theme 1: There is limited consensus of the Advisory Panel on whether there is a hierarchy among the various accountability applications

The Advisory Panel considered at length the question of whether there is a hierarchy among the various accountability applications (e.g. pay-for-performance, public reporting). Several members of the Panel argued that a hierarchy among the various accountability applications could be determined. The hierarchy would be based on the potential risk of misclassifying providers in the program and the subsequent financial impact to providers. Some Panel members noted that measures used for pay for performance programs, particularly programs that are designed as penalty programs, should have the highest level of requirements for measure testing and evidence. There is a considerable impact to the health care system if measures used in payment programs inappropriately misclassify providers and take away resources that are needed to support performance improvement.

Other members of the Panel argued that most accountability applications have a financial impact on providers, regardless of whether they have direct financial provider penalties, and thus a hierarchy of financial impact to providers would be difficult to determine. Pay for performance programs, but also public reporting, health plan network design, and other programs, have the potential to have a financial impact on providers. The use of performance measures can also have a financial impact on providers by informing consumer choice on the selection of providers and directing patients away from low-performing providers.

Others argued that a hierarchy based on financial impact to providers may be misguided, and that any hierarchy is inherently dependent on the perspective of the user. This risk of misclassification of providers also has a financial, and quality of care, impact for patients, consumers, and purchasers. These

panel members noted that the scientific rigor used to examine the risk of misclassification of providers is required in multiple accountability applications to effectively drive purchasing and selection decisions by purchasers and consumers. Measures that are endorsed by NQF should produce reliable and valid information, regardless of the intended use of the measure, and make the level of scientific rigor transparent for those selecting measures for implementation in programs. Finally, several panel members noted that there is a lack of evidence that any one of these types of programs has a lower risk of misclassification, greater impact on provider behavior, and thus systematically improves outcomes more than any other type of program.

Theme 2: There is a qualitative difference between measures used for quality improvement (QI) only and those used in other applications.

Measures for QI only are used by providers to evaluate their performance over time or among a group of providers and are not typically used publicly for comparative purposes. As such, the panel noted that measures for internal quality improvement may require a different standard for evidence and testing. There are a myriad of organizations developing measures for use in their QI programs and different criteria for QI-only measures may help to create an environment where these organizations can more easily use and share measures.

Theme 3: Field experience could be a useful addition in the evaluation of measures used in accountability applications.

The current NQF endorsement criteria require empirical reliability testing on a representative sample of patients, and a minimum of face validity testing. The Panel noted that this is a minimum acceptable standard for evaluating the scientific properties of a measure; however, additional data from field experience of measures would be useful. These data could include information on the gap identified from the measure as specified, or an assessment of how well the measure identifies variation and meaningful differences of performance across providers. The Panel acknowledged that there is no consensus definition of what field experience would include but they noted that it is important to create a mechanism to monitor how a measure is performing once it is implemented and to incentivize measure developers to provide these data for review.

Recommendations

The recommendations that emerged from the Advisory Panel deliberations are outlined below.

Recommendation 1: Move forward with grading measures based on the NQF endorsement criteria

The Advisory Panel generally agreed that NQF should begin to grade measures based on how well they meet the NQF endorsement criteria. By using grades, NQF can apply the same criteria to all measures, but make transparent how well the measures met the criteria and potentially how well the measure performs when it is implemented. Endorsement would change from a simple yes/no to a categorical grading scale. When a measure meets the criteria to a greater degree, then it would receive a higher grade. The grade should not be directly linked to a specific use.

- **AAA**: Highest grade measures: AAA measures received the highest rating on all of the must-pass endorsement criteria, specifically, evidence, opportunity for improvement, empirical validity testing, and reliability testing at the measure score level. Measures should have a high or moderate rating on feasibility, and usability and use. These measures will have demonstrated field experience for at least one year with assessments of the gap in measure performance, or an assessment of how well the measure identifies variation and meaningful differences of performance across providers.
- **AA**: High grade measures: AA measures received moderate or high ratings for all NQF criteria. AA measures have moderate or high ratings for evidence, opportunity for improvement, empirical validity testing, reliability testing at the data element or measure score level, feasibility, and usability and use. In contrast to AAA measures, there may be limited field experience with performance data.
- **A**: Moderate grade measures: A measures received moderate ratings for all NQF criteria for endorsement with limited field experience and performance data.

The Advisory Panel considered two other potential proposals. One proposal would directly link the measure grades to specific uses of measures. For example, AAA measures would be recommended as appropriate for pay-for-performance programs with downside financial risk (e.g. penalty programs), AA measures would be appropriate for use in pay-for-reporting with upside or downside financial risk and for public reporting, and A measures would be appropriate for all other uses. Generally, the Advisory Panel did not support this proposal because there was limited agreement on the hierarchy of specific measure uses to allow for the linking of uses to ratings. The Panel also considered whether the Measures Applications Partnership (MAP) should interpret the grades and how they can be applied to individual programs. There was general agreement that, while the proposal was promising, the MAP process should define how, or even if, the grades inform their selection of measures for programs.

Recommendation 2: Incorporate "field experience" in the measure endorsement process

While the Advisory Panel struggled to precisely define "field experience," there was consensus that it would include the ability to review performance data on the performance measure prior to widespread implementation. Ideally, measure implementers would allow those being measured to have an opportunity to review their performance results. Measure developers should have an audit procedure to ensure that any implementation challenges are identified and shared with end-users. Field experience of measures could include analysis of the measure in the target population to better assess the performance of the measure. This analysis could include assessments of the gap in measure performance, an assessment of how well the measure identifies variation and meaningful differences of performance across providers, and an assessment of the benefits and risks associated with the measure by end-users prior to widespread implementation. The Panel noted that the requirement for field experience would likely evolve with the maturity of the measure.

Recommendation 3: Consider the rapidly evolving uses of measures and incentivize grade advancement through further use and testing

The Panel discussed a different approach for measure maintenance, in which experience with measures is tracked over time. The newly approved changes in NQF measure maintenance process already emphasize the importance of measure use and experience over re-submission of evidence and measure testing results. The Panel reaffirmed the importance of identifying positive and negative unintended consequences when a measure is in use, though recognized the difficulty of accessing those data by measure developers. Given the recommended shift to endorsement grades, there was a recommendation that the measure maintenance process be utilized as a collaborative opportunity to raise the measure's grade through further use and testing. Measure maintenance could create a positive incentive for measure developers to perform additional analyses or recommended updates on measures to move up the grading scale. This could create a more collaborative opportunity between developers and end-users to access the data required to provide updates that meet the needs of a rapidly changing measurement enterprise.

Recommendation 4: Consider different endorsement standards for QI measures

Given the distinctions between QI and other uses of measures, the Panel encouraged the development of a NQF process that would allow QI-only measures to be evaluated differently than other measures, and, potentially, approved. While QI-measures should demonstrate opportunity for improvement or variation across providers, the review process could potentially lower requirements for the must-pass requirements of evidence, reliability, and validity, compared to the current endorsement process. Specifically, the Panel noted that QI-only measures could potentially be approved with lower levels of evidence (e.g., no clear process-outcome linkage). Further, measures for QI may require only face validity with recognition that reliability may emerge from use. The Panel questioned the perceived need for a NQF evaluation for QI measures from end-users. Some panel members emphasized the importance of NQF review of QI measures to avoid re-invention by end-users (including providers, plans and States), and the importance of driving toward greater standardization of QI measures, especially those used in national efforts. Since this would require a new process for NQF, the Panel specifically seeks comment on the usefulness of an approval process for QI measures.

Recommendation 5: Encourage the Measures Applications Partnership (MAP) to consider how measure grades can be used in the selection of measures for programs

The Advisory Panel encourages the Measures Applications Partnership (MAP) to consider how the grades can be used when selecting individual measures for programs. For example, in an effort to align program and measure attributes, the MAP may determine that an individual program requires only AAA measures, where as another program may only need AA measures. The Advisory Panel generally agreed that the MAP Coordinating Committee would be most appropriate to consider this approach in future work.

Recommendation 6: Pursue future work to consider the interaction between program attributes and individual measure attributes

The Advisory Panel urged that future work should define key measure attributes and program attributes, examine their interaction, and give program implementers guidance on which measures may be better suited for implementation in specific programs. The Panel identified a preliminary set of measure attributes that can be considered, including; (1) inclusion and exclusion criteria, (2) potential for misclassification based on reliability and validity testing results, and (3) the precision of the risk adjustment models. Further, a set of program attributes may include; (1) methods used to define performance categories (e.g. measure score thresholds), (2) whether or not statistical tests are used to distinguish between performance categories and the approach to those tests, and (3) the nature of the financial incentive (e.g. tied to performance or improvement).

This work can begin with categorizing the measure and program attributes, and move further to provide guidance on the interaction of these various elements would advance health care measurement science by identifying more precisely the likelihood that a measure will perform well in the context of a given program design. An examination of measure attributes and program attributes will reveal other key principles for the endorsement and application of measures. These key principles can help to create a test environment in which measures and programs may be more precisely matched to drive health system performance improvement.

Request for Comment

The Advisory Panel seeks public comment on the key themes and the recommendations that emerged from its deliberations of the charge outlined by the NQF Board. The Panel will consider these comments during an upcoming comment call on October 20, 2015. Following this call, the recommendations will be finalized and presented to the Consensus Standards Approval Committee (CSAC) and the NQF Board.

Appendix A: Panel Members and NQF Staff

PANEL MEMBERS

Helen Darling, MA
National Business Group on Health
Chair, National Quality Forum Board of Directors
Washington, DC

Elizabeth E. Drye, MD, SM
Yale School of Medicine, Center for Outcomes Research & Evaluation
New Haven, CT

Lee A. Fleisher, MD
Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA

Don Goldmann, MD
Institute for Healthcare Improvement
Cambridge, MA

Kate Goodrich, MD, MHS
Centers for Medicare & Medicaid Services
Washington, DC

Bruce L. Hall, MD, MBA, PhD
BJC Healthcare
Saint Louis, MO

Mary Beth Landrum, PhD
Harvard Medical School
Boston, MA

Beth A. McGlynn, PhD, MPP
Kaiser Permanente Center for Effectiveness & Safety Research
Pasadena, CA

Jonathan Perlin, MD, MSHA, PhD, FACP
Hospital Corporation of America
Nashville, TN

Andrew Ryan, PhD

University of Michigan Center for Health Outcomes and Research
Ann Arbor, MI

Cristie Upshaw Travis, MSHA
Memphis Business Group on Health
Memphis, TN

Lina Walker, PhD
AARP Public Policy Institute
Washington, DC

Rachel Werner, MD, PhD
University of Pennsylvania
Philadelphia, PA

NQF STAFF

Helen Burstin, MD, MPH
Chief Scientific Officer

Taroon Amin, PhD, MPH
Consultant

Suzanne Theberge, MPH
Senior Project Manager

Poonam Bal, MSHA
Project Manager

Kaitlynn Robinson-Ector, MPH
Project Analyst

National Quality Forum
1030 15th St NW, Suite 800
Washington, DC 20005
http://www.qualityforum.org

ISBN [DRAFT]