NATIONAL QUALITY FORUM
Driving measurable health improvements together

Cost and Efficiency, Spring 2022 Cycle: CDP Report

TECHNICAL REPORT FOR COMMENT
JANUARY 30, 2023

This report is funded by the Centers for Medicare & Medicaid Services under contract HHSM-500-2017-00060I Task Order HHSM-500-T0001.

https://www.qualityforum.org

Contents

Executive Summary

As healthcare expenditures continue to grow, it is crucial to understand how resources are utilized to maximize quality in the healthcare system. Healthcare cost measurement continues to be a critical component in assessing the United States (U.S.) healthcare system. Measures in the Cost and Efficiency portfolio are essential to evaluate the efficiency of care and improve value through changes in practice. Improving U.S. health system efficiency can simultaneously reduce cost growth and improve the quality of care provided. National Quality Forum's (NQF) Cost and Efficiency Standing Committee oversees NQF's portfolio of cost and resource use measures, which includes both condition-specific and non–condition-specific measures. This portfolio contains 13 measures: seven condition-specific measures and six non–condition-specific measures.

For this cycle, the Standing Committee evaluated three newly submitted measures against NQF's standard evaluation criteria. The Standing Committee recommended all three measures for endorsement, and the Consensus Standards Approval Committee (CSAC) upheld the Standing Committee's recommendations.

The Standing Committee endorsed the following measures:

- **NQF #3623** Elective Primary Hip Arthroplasty Measure (Centers for Medicare & Medicaid Services [CMS]/Acumen, LLC)
- **NQF #3625** Non-Emergent Coronary Artery Bypass Graft (CABG) Measure (CMS/Acumen, LLC)
- **NQF #3626** Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure (CMS/Acumen, LLC)

Brief summaries of the measures and their evaluations are included in the body of the report; detailed summaries of the Standing Committee's discussion and ratings of the criteria for each measure are in Appendix A.

Introduction

United States (U.S.) healthcare spending was estimated to have reached $4.3 trillion and projected to have grown by 4.2 percent in 2021.[1] This growth is projected to continue increasing to $6.8 trillion by 2030.[1] Currently, U.S. healthcare costs are growing 1.1 percent faster than the annual gross domestic product (GDP), and it is estimated that U.S. healthcare spending will account for almost 20 percent of the GDP by 2028.[2] Medicare is expected to experience the fastest spending growth as a result of having the highest projected enrollment growth (7.6 percent per year over 2019–2028).[2] U.S. hospital spending and physician and clinical service spending rates are also expected to increase in 2022 (6.9 percent for hospital spending and 6.2 percent for physician and clinical service spending).[1]

In addition to increasing healthcare costs, the U.S. spends far more on healthcare when compared to similar high-income countries yet has worse health outcomes, including a lower life expectancy and higher chronic disease burdens, obesity rates, hospitalizations from preventable causes, and rates of

avoidable deaths.[3] Benchmarking the spending and performance of the healthcare system is essential to assessing and improving the efficiency and value of the U.S. healthcare system. Furthermore, healthcare cost measurement is necessary to improve the quality of care that is provided to consumers.

The Cost and Efficiency Standing Committee reviewed three measures during this spring 2022 measure evaluation cycle. These measures focused on elective primary hip arthroplasty (NQF #3623), nonemergency coronary artery bypass graft (CABG) (NQF #3625), and lumbar spine fusion for degenerative disease (NQF #3626). The Scientific Methods Panel (SMP) originally reviewed all three measures in spring 2021. Due to capacity issues that emerged during the spring 2021 cycle, as a result of COVID-19 and competing priorities, the endorsement review of these measures was deferred to the spring 2022 review cycle.

NQF Portfolio of Performance Measures for Cost and Efficiency Conditions

The Cost and Efficiency Standing Committee (Appendix C) oversees NQF's portfolio of cost and resource use measures (Appendix B), which includes both condition-specific and non–condition-specific measures. This portfolio contains 16 measures: 10 condition-specific measures and six non–condition-specific measures.

Cost and Efficiency Measure Evaluation

On July 12, 2022, the Cost and Efficiency Standing Committee evaluated three new measures undergoing endorsement review against NQF's standard measure evaluation criteria.

Table 1. Cost and Efficiency Measure Evaluation Summary

Measure	Maintenance	New	Total
Measures under review for endorsement	0	3	3
Measures endorsed	0	3	3

Scientific Methods Panel Measure Evaluation

Prior to the Standing Committee's review, the SMP reviewed all three measures for this topic area during the spring 2021 cycle. These measures were moved to the spring 2022 cycle due to capacity issues that emerged during the spring 2021 cycle. The SMP passed all three measures on reliability and validity during its measure evaluation.

A meeting summary detailing the SMP's measure evaluation for the spring 2021 cycle is available on the SMP webpage.

Comments Received Prior to Standing Committee Evaluation

NQF accepts comments on endorsed measures on an ongoing basis through the Quality Positioning System (QPS). In addition, NQF solicits comments for a continuous period during each evaluation cycle via an online tool located on the project webpage. For this evaluation cycle, the commenting period

opened on May 18, 2022, and pre-meeting commenting closed on June 15, 2022. Prior to June 15, 2022, three comments were submitted and shared with the Standing Committee prior to the measure evaluation meeting (Appendix F).

Comments Received After Standing Committee Evaluation

The continuous public commenting period with NQF member support closed on September 26, 2022. Following the Standing Committee's evaluation of the measures under review, NQF received three comments from one organization, which is an NQF member organization, pertaining to the draft report and the measures under review (Appendix G). All comments for each measure under review have also been summarized in Appendix A.

NQF members had the opportunity to express their support ("support" or "do not support") for each measure submitted for endorsement consideration to inform the Standing Committee's recommendations during the commenting period. One NQF member expressed "do not support" for NQF #3623, NQF #3625, and NQF #3626.

Overarching Themes

During the Standing Committee's discussion of the measures, several overarching issues emerged that were factored into the Standing Committee's ratings and recommendations for multiple measures and are not repeated in detail with each individual measure.

Linking Cost and Quality Measures

During the spring 2022 measure evaluation proceedings, the Standing Committee questioned whether the developer was able to demonstrate that the hospitals being measured could demonstrate improvements in costs while ensuring similar or higher levels of quality. Specifically, the Standing Committee was interested in the relationship between performances on cost and related quality measures. Some Standing Committee members expressed concern with the unintended consequence of performing well on cost measures at the expense of lower quality performance. While the developer did report that they performed some analysis in response to this question, it is not currently requested or required as part of the NQF submission process.

Social Risk Adjustment

While some of the measures this cycle did test for social risk factors (SRFs) (e.g., age, race, ethnicity, gender, social relationships, and geographic location) for the measure's risk adjustment model, namely dual eligibility, some of the measures under review did not include these SRFs in the final model. The Standing Committee recognized the need to ensure that providers who serve people with SRFs are not penalized unfairly due to a lack of social risk adjustment. While the Standing Committee did note that it is important to maximize the predictive value of a risk adjustment model, understanding the role SRFs play in clinical-cost episodes is critical. The impact of SRFs on cost and resource measures is unique because these factors may ultimately increase overall costs through poor transitions and hand-offs or potentially lower resource use due to access-to-care challenges. Each cost measure should be examined case by case to understand the role of patient SRFs in the measure.

The Standing Committee asked NQF staff whether any work is currently being done at NQF to address the concerns regarding SRF adjustment within quality measurement. In response, NQF staff stated that NQF is currently developing technical guidance for social and/or functional status-related risk adjustment within quality measurement. This guidance will help to evolve NQF's current criteria, which will occur after 2022. Therefore, the Standing Committee must review the measures for the spring 2022 cycle under NQF's current measure evaluation criteria.

Summary of Measure Evaluation

The following brief summaries of the measure evaluation highlight the major issues that the Standing Committee considered. Details of the Standing Committee's discussion and ratings of the criteria for each measure are included in Appendix A.

NQF #3623 Elective Primary Hip Arthroplasty Measure (Centers for Medicare & Medicaid Services [CMS]/Acumen, LLC): Endorsed

Description: The Elective Primary Hip Arthroplasty episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who receive an elective primary hip arthroplasty during the performance period. The measure score is a clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from the 30 days prior to the clinical event that opens or "triggers" the episode, through 90 days after the trigger. Patient populations eligible for the Elective Primary Hip Arthroplasty measure include Medicare beneficiaries enrolled in Medicare Parts A and B; **Level of Analysis:** Clinician: Individual, Clinician: Group/Practice; **Setting of Care:** Ambulatory Care: Clinician Office, Other, Inpatient/Hospital, Ambulatory Care: Clinic/Urgent Care; **Type of Measure**: Cost and Resource use; **Data Source**: Claims

This group/practice- and individual clinician-level measure was newly submitted for endorsement. It is publicly reported in the Quality Payment Program (QPP) Merit-Based Incentive Payment System (MIPS).

The Standing Committee reviewed the data the developer provided, which demonstrated a high prevalence of total hip arthroplasties representing 0.8 percent for the general population and increasing with age to 1.5 percent at 60 years of age and 5.9 percent by 90 years of age. During the discussion on opportunities for improvement, the Standing Committee noted that the performance gap data indicated a mean score of 1.03 (standard deviation [SD] of 0.12, interquartile range [IQR] of 0.15) at the clinician-group level and a mean score of 1.00 (SD of 0.12, IQR of 0.15) at the individual-clinician level. The Standing Committee agreed that the IQR of 0.15 would have translated into significant overall cost savings for Medicare if the performance on this measure had moved from the 75th to the 25th percentile of cost. The Standing Committee also cautioned that while a performance gap in spending was present, it was difficult to ascertain the actions clinicians can take to impact this variation and how it relates to overall patient care quality. The Standing Committee ultimately passed the measure on high impact and improvement opportunities.

The Standing Committee noted that the SMP previously reviewed this measure in spring 2021 and passed it with a rating of high on reliability and a rating of moderate on validity. The Standing Committee agreed with the SMP's evaluation, which stated that the developer's signal-to-noise ratio (SNR) and split-sample reliability testing were sufficient, and the testing results indicated a robust

measure of score reliability. The Standing Committee noted that the developer conducted empirical and face validity testing at the accountable-entity level. During the discussion on validity, the Standing Committee raised several concerns, specifically with the small size of the developer's initial technical expert panel (TEP), the correlation of this measure with a similar NQF-endorsed resource measure instead of a quality measure, and the merits of the attribution and shared accountability for measure performance (i.e., primary and assisting surgeon). The developer explained that the subsequent TEP was more significant (n=29 members) and included experts in musculoskeletal disease management with affiliations in 26 organizations and specialty societies. Regarding the attribution approach, the developer noted that the primary and assisting surgeons are both attributed because they have joint responsibility for the cost measure. Addressing the Standing Committee's concern with the quality and cost correlation, the developer noted that in addition to the correlation analysis performed with NQF #2158 *Medicare Spending per Beneficiary (MSPB) Hospital Measure*, they performed correlation analysis with NQF #3495 *Hospital-Wide 30-Day, All-Cause, Unplanned Readmission Rate (HWR) for the Merit-Based Incentive Payment System (MIPS)-Eligible Clinician Groups*. During the discussion, the developer reported a Pearson correlation of 0.27 amongst providers with lower costs and complication rates that they considered a medium correlation.

The Standing Committee also expressed concern with the lack of social risk adjustment in the risk model. The developer explained that the measurement results were stratified by dual-eligibility status and found that the risk-adjusted cost for both dual-eligible and non–dual-eligible episodes increases among providers with higher dual-eligible populations (i.e., providers with higher dual-eligible beneficiaries may perform worse). The developer expressed concern that risk-adjusting for dual status could unintentionally remove some of the difference in performance due to the provider-level effect versus the individual-level effect. Ultimately, the Standing Committee accepted the developer's responses to the concerns raised, agreed with the SMP, and passed the measure on validity.

The Standing Committee agreed that the measure is feasible and that the data elements required for the measure are readily available and could be captured without undue burden. The Standing Committee also acknowledged that this is a new measure but that the developer did not provide any improvement data. The Standing Committee questioned how clinicians could improve the quality of care while reducing cost when healthcare settings and services are determined by healthcare systems where physicians are employed. The developer explained that clinicians receive field reports containing cost performance categories that can be further broken down into specific services and settings to identify areas of improvement. The Standing Committee accepted the developer's response and passed the measure on feasibility, use, usability, and overall suitability for endorsement.

During the post-evaluation commenting period, one comment was received. This comment did not express support for the measure, noting the lack of correlations of the cost measure with quality measures and the omission of social risk factors in the risk adjustment model. Specifically, the commenter suggested that testing should demonstrate use of the measure in MIPS to make meaningful distinctions in costs associated with the care provided to the patients. The developer responded, noting the measure has high reliability at both the TIN [0.86] and TIN-NPI [0.80] levels. Furthermore, the developer addressed the commenter’s concern by stating that the threshold set by CMS through regulatory processes is pertinent to any evaluation of reliability and noted the SMP passed the measure on the reliability criterion. The Standing Committee recommended the measure for initial endorsement.

No related and competing measures were identified for this measure. The CSAC upheld the Standing Committee's decision to recommend the measure for endorsement. No appeals were received.

NQF #3625 Non-Emergent Coronary Artery Bypass Graft (CABG) Measure (CMS/Acumen, LLC): Endorsed

Description: The Non-Emergent CABG episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo a CABG procedure during the performance period. The measure score is the clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from 30 days prior to the clinical event that opens, or "triggers," the episode through 90 days after the trigger. Patient populations eligible for the Non-Emergent CABG measure include Medicare beneficiaries enrolled in Medicare Parts A and B; **Level of Analysis:** Clinician: Individual, Clinician: Group/Practice; **Setting of Care:** Inpatient/Hospital; **Type of Measure**: Cost and Resource use; **Data Source**: Claims

This group/practice- and individual clinician-level measure was newly submitted for endorsement. It is publicly reported in the QPP MIPS program.

While the Standing Committee did acknowledge a high prevalence of nonemergent CABG surgeries among Medicare beneficiaries reflecting substantial Medicare expenditures, it noted an overall downward trajectory and steady decline of CABG cases and mortality. The developer explained that with the advancements in interventional cardiology, the number of CABG procedures would continue to decrease. The Standing Committee cautioned that while a performance gap in spending was present, it is difficult to ascertain the actions clinicians can take to impact this variation and how it relates to overall patient care quality. One Standing Committee member noted that the measure aims to reduce the number of avoidable readmissions and appropriate post-acute care and questioned the rationale for making this measure a cost measure instead of a quality measure. The developer explained that the MIPS cost performance category requires measures based on care episode groups. The developer further noted that they selected the CABG episode because it is a high-frequency, high-cost care area. The Standing Committee accepted the developer's rationale and passed the measure on high impact and opportunity for improvement.

The Standing Committee noted that the SMP previously reviewed and passed this measure in spring 2021 with a rating of moderate on both reliability and validity. While the Standing Committee agreed that the reliability testing was robust, one Standing Committee member requested clarification on why the developer selected the 10-episode case minimum. The developer explained that careful consideration was given to both coverage and reliability when determining the case minimum to ensure that smaller providers with lower case volumes are assessed. The Standing Committee agreed with the SMP that the reliability testing was appropriate and that the testing results indicated moderate measure score reliability.

The Standing Committee reviewed the validity testing the developer conducted at the performance measure score level. While the Standing Committee agreed that the validity testing was robust, it raised concerns about the high number of exclusions. The developer explained that the exclusion logic is designed to capture only nonemergent CABG procedures and the exclusions selected to ensure the

measure is not accidentally capturing emergent procedures. The Standing Committee accepted the developer's rationale, agreed that the validity testing was sufficient, and passed the measure on validity.

The Standing Committee agreed that the measure is feasible and that the data elements required for the measure are readily available and could be captured without undue burden. While the Standing Committee did acknowledge that this is a new measure and that the developer did not provide any improvement data, it raised concerns about how the measure's performance results can be used to improve care further. Specifically, the Standing Committee questioned how the developer plans to differentiate between natural variation and areas of actual improvement in care. The developer will continue to monitor the impact of the measure and noted that they expect an early reduction in cost to occur and then a gradual flattening out and convergence across providers. The Standing Committee ultimately passed the measure on feasibility and use.

During the discussion of unintended consequences, one Standing Committee member noted that opportunities for significant cost savings might be excluded when outlier cases are eliminated from the data, as this may be where the actual waste and inefficiencies reside. The developer clarified that 1 percent of episodes at both ends of the distribution are excluded in the areas for which the risk adjustment model cannot predict cost accurately. The Standing Committee appreciated the developer's response and ultimately passed the measure on usability and overall suitability for endorsement.

During the post-evaluation commenting period, one comment was received. This comment did not express support for the measure, noting the lack of correlations of the cost measure with quality measures and the omission of social risk factors in the risk adjustment model. Specifically, the commenter suggested that testing should demonstrate use of the measure in MIPS to make meaningful distinctions in costs associated with the care provided to the patients. The developer responded, noting the measure has high reliability at both the TIN [0.84] and TIN-NPI [0.75]) levels. Furthermore, the developer addressed the commenter's concern by stating that the threshold set by CMS through regulatory processes is pertinent to any evaluation of reliability and noted the SMP passed the measure on the reliability criterion. The Standing Committee recommended the measure for initial endorsement. No related and competing measures were identified for this measure. The CSAC upheld the Standing Committee's decision to recommend the measure for endorsement. No appeals were received.

NQF #3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure (CMS/Acumen, LLC): Endorsed

Description: The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo surgery for lumbar spine fusion during the performance period. The measure score is the clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from 30 days prior to the clinical event that opens, or "triggers," the episode through 90 days after the trigger. Patient populations eligible for Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure include Medicare beneficiaries enrolled in Medicare Parts A and B; **Level of Analysis:** Clinician: Group/Practice, Clinician: Individual; **Setting of Care:** Inpatient/Hospital, Other, Ambulatory Care: Clinic/Urgent Care; **Type of Measure**: Cost and Resource use; **Data Source**: Claims

This group/practice- and individual clinician-level measure was newly submitted for endorsement. It is reported publicly in the QPP MIPS program.

The Standing Committee reviewed data demonstrating a high prevalence of degenerative lumber conditions affecting more than 6 million Medicare patients and a total admission expenditure for lumbar spine fusion surgeries exceeding $3.6 billion in 2013. During the discussion on opportunities for improvement, a Standing Committee member questioned what services tend to drive cost-per-case variability. The developer explained that acute readmissions and post-acute care have the most influence on cost. The Standing Committee agreed that this measure captures an area of high-impact and resource use that warrants a national performance measure and passed the measure on both criteria.

The Standing Committee noted that the SMP reviewed this measure in spring 2021 and passed it with a rating of moderate on reliability and validity. The Standing Committee agreed with the SMP's evaluation, which stated that the developer's SNR and split-sample reliability testing were sufficient and that the testing results indicated moderate measure score reliability. While the Standing Committee agreed that the validity results were robust, a Standing Committee member requested further clarification on how the developer applied the model across the three subgroups (i.e., one subgroup for the three distinct levels of procedures). The developer explained that they stratified all episodes into three mutually exclusive subgroups and applied the risk adjustment model separately within each of the three subgroups. The developer further explained that the three subgroup scores are rolled up at the provider level to calculate the overall measure score. One Standing Committee member noted that base and race data are challenging to parse out from the current risk model, which combines three components (i.e., base, dual-eligibility status, and race). The Standing Committee member suggested that the developer consider a risk model that only provides a base population plus race. The Standing Committee ultimately passed the measure on validity.

The Standing Committee agreed that the measure is feasible and that the data elements required for the measure are readily available and could be captured without undue burden. During the discussion of usability, the Standing Committee raised concern about the potential for undertreatment and the unintended consequences of pain management and opioid prescribing among patients undergoing lumbar spine procedures. The developer explained that the cost drivers are related to adverse outcomes; undertreatment typically results in costly adverse events that the measure will capture within the 90-day postoperative period. The developer further noted that drugs are included in the service assignment and highlighted the importance of opioid use quality measures, which look specifically at prescribing practices and use. The Standing Committee accepted the developer's response and passed the measure on feasibility, use, usability, and overall suitability for endorsement.

During the post-evaluation commenting period, one comment was received. The comment did not express support for the measure, noting the lack of correlations of the cost measure with quality measures and the omission of social risk factors in the risk adjustment model. Specifically, the commenter suggested that testing should demonstrate use of the measure in MIPS to make meaningful distinctions in costs associated with the care provided to the patients. The developer responded, noting the measure has high reliability at both the TIN [0.78] and TIN-NPI [0.72]) levels. Furthermore, the developer addressed the commenter's concern by stating that the threshold set by CMS through

regulatory processes is pertinent to any evaluation of reliability and noted the SMP passed the measure on the reliability criterion. The Standing Committee recommended the measure for initial endorsement. No related and competing measures were identified for this measure. The CSAC upheld the Standing Committee's decision to recommend the measure for endorsement. No appeals were received.

References

1 Poisal JA, Sisko AM, Cuckler GA, et al. National Health Expenditure Projections, 2021–30: Growth To Moderate As COVID-19 Impacts Wane. *Health Affairs*. 2022;41(4):474-486. https://www.healthaffairs.org/doi/abs/10.1377/hlthaff.2022.00113. Last accessed August 2022.

2 Centers for Medicare & Medicaid Services. NHE Fact Sheet | CMS. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet. Last accessed August 2022.

3 The Commonwealth Fund. U.S. Health Care from a Global Perspective, 2019 | Commonwealth Fund. https://www.commonwealthfund.org/publications/issue-briefs/2020/jan/us-health-care-global-perspective-2019. Last accessed August 2022.

Appendix A: Details of Measure Evaluation

Rating Scale: H=High; **M**=Moderate; **L**=Low; **I**=Insufficient; **NA**=Not Applicable

NQF ensures that quorum is maintained for all live voting. Quorum is 66 percent of active Standing Committee members minus any recused Standing Committee members. Due to the exclusion of recused Standing Committee members from the quorum calculation, the required quorum for live voting may vary among measures. During the meeting, the quorum required for voting was not achieved (10 out of 15 active Standing Committee members). Therefore, the Standing Committee discussed all criteria for each measure and voted after the meeting using an online voting tool. No voting occurred during the post-comment call on October 27, 2022. The Standing Committee received a recording of the meeting and a link to submit online votes. Voting results are provided below.

A measure is recommended for endorsement by the Standing Committee when greater than 60 percent of voting members select a passing vote option (i.e., Pass, High and Moderate, or Yes) on all must-pass criteria and overall suitability for endorsement. A measure is not recommended for endorsement when less than 40 percent of voting members select a passing vote option on any must-pass criterion or overall suitability for endorsement.

Measures Endorsed

NQF #3623 Elective Primary Hip Arthroplasty Measure (CMS/Acumen, LLC)

Measure Worksheet | Specifications

Description: The Elective Primary Hip Arthroplasty episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who receive an elective primary hip arthroplasty during the performance period. The measure score is a clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from the 30 days prior to the clinical event that opens or "triggers" the episode, through 90 days after the trigger. Patient populations eligible for the Elective Primary Hip Arthroplasty measure include Medicare beneficiaries enrolled in Medicare Parts A and B.

Numerator Statement: Not required for cost measures.

Denominator Statement: Not required for cost measures.

Exclusions: Exclusions are used in the Hip Arthroplasty Measure to ensure a homogenous and comparable patient population within the measure's focus on elective primary hip arthroplasties. These exclusions focus on removing patients where fair comparisons cannot be made across providers, preventing potential threats to measure validity and ensuring that episodes provide meaningful information to attributed clinicians. These exclusions are listed below:

- Episodes where patient death date occurred before the episode end.
- Episodes where the trigger claim was not in an ambulatory/office-based care setting, IP hospital, OP hospital, or ASC based on its place of service.
- Episodes with inpatient procedures, where the inpatient stay did not occur in either an acute hospital as defined by subsection (d) or in an acute hospital in Maryland.
- Episodes with inpatient procedures, where the inpatient stay did not have a relevant MS-DRG code.
- Episodes in which the patient underwent a staged or same-day bilateral hip arthroplasty.
- Episodes where the hip replacement was performed due to cancer, hip fracture, or trauma.
- Episodes where the patient had a congenital deformity of the hip, osteomyelitis of the hip or femur, or a septic joint.
- Episodes classified as outlier cases.

Further explanation and rationale for each of the measure exclusions above can be found in Section S.9.1 of the Intent to Submit form. Please also see Section 2b6 (Missing Data Analysis and Minimizing Bias) of this testing form for more information on exclusions implemented as part of data processing.

Given the rationale for the exclusions noted above, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a higher or lower mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). To demonstrate this, we examined the distributions of observed cost and ratio of observed over expected spending (calculated by applying existing risk factor coefficients to the excluded episodes) for each excluded population. We then compared the cost characteristics of the excluded episodes to that of episodes included in the measure to assess the distinctness between the two patient cohorts.

Adjustment/Stratification: Statistical Risk Model

Level of Analysis: Clinician: Group/Practice, Clinician: Individual

Setting of Care: Ambulatory Care: Clinician Office, Other, Ambulatory Care: Clinic/Urgent Care, Inpatient/Hospital

Type of Measure: Cost/Resource Use

Data Source: Claims

Measure Steward: Centers for Medicare & Medicaid Services (CMS)

STANDING COMMITTEE MEETING July 12, 2022

1. Importance to Measure and Report:

(1a. High Impact, 1b. Opportunity for Improvement)

1a. High Impact and 1b. Opportunity for Improvement: **Total votes- 10; H-2; M-8; L-0; I-0**

Rationale:

- The Standing Committee reviewed data demonstrating a high prevalence of total hip arthroplasties representing 0.8 percent for the general population and increasing with age to 1.5 percent at 60 years of age and 5.9 percent by 90 years of age.
- The Standing Committee acknowledged that the demand for hip arthroplasties is anticipated to double between 2005 and 2030, increasing from an estimated 2.5 million patients with hip arthroplasties in 2010.
- The Standing Committee agreed that this measure captures an area of high impact and resource use that warrants a national performance measure.
- During the discussion of the performance gap, the Standing Committee noted that the performance gap data indicated a mean score of 1.03 (SD of 0.12, IQR of 0.15) at the clinician-group level and a mean score of 1.00 (SD of 0.12, IQR of 0.15) at the individual-clinician level.
- The Standing Committee agreed that the IQR of 0.15 would have translated into significant overall cost savings for Medicare if the performance on this measure had moved from the 75th to the 25th percentile of cost.
- The Standing Committee cautioned that while a performance gap in spending was present, it was difficult to ascertain the actions clinicians can take to impact this variation and how it relates to overall patient care quality.
- The Standing Committee ultimately passed the measure on the opportunity for improvement.

2. Scientific Acceptability of Measure Properties:

(2a. Reliability - precise specifications, testing; 2b. Validity - testing, threats to validity)

2a. Reliability: **Total votes-10; H-4; M-6; L-0; I-0**; 2b. Validity: **Total votes-10; H-0; M-8; L-1; I-1**

Rationale:

- The SMP reviewed this measure and passed it with a rating of high on reliability **(Total votes-8; H-7, M-1, L-0, I-0**) and a rating of moderate on validity **(Total votes-7; H-0, M-5, L-2, I-0).**
- The Standing Committee agreed with the SMP's evaluation, which stated that the developer's SNR and split-sample reliability testing were appropriate and that the testing results indicated a robust measure of score reliability.
- Furthermore, several Standing Committee members expressed caution with correlating the measure with another NQF-endorsed resource use measure (NQF #2158 *Medicare Spending per Beneficiary* [MSPB] *Hospital Measure*) due to its limitations, specifically because it analyzes two resource use measures without an external construct.

- The developer responded by explaining that a correlation analysis was performed using the MIPS risk-standardized complication rate for hip arthroplasty measure numerator and reported a Pearson correlation of 0.27 amongst providers with lower costs and complication rates. Several Standing Committee members pointed out that this was a low correlation.
- The developer noted that correlating cost measures with MIPS quality measures comes with challenges, specifically the small sample sizes, minimum data completion requirements (i.e., 70 percent of data for eligible beneficiaries in the denominator), and the potential for selection bias (i.e., reporting clinicians select highest scoring measures for submission).
- Lastly, the developer noted there is a potential to explore correlation with a new claims-based risk-standardized complication rate measure included in the 2021 MIPS program when the data become available at the end of June 2022.
- The Standing Committee discussed the measure's scope of improvement, emphasizing that lower cost does not necessarily correlate to improved quality. The Standing Committee noted that if cost and quality are not highly correlated, then there may be a risk of potential unintended consequences, such as lowering the quality of care provided.
- The Standing Committee also discussed the merits of attribution to both the primary clinician and the assisting clinicians. The developer clarified that each clinician has joint responsibility in terms of measurement in the cost measure.
- The Standing Committee expressed concern with the lack of social risk adjustment in the measure's statistical risk model. The developer explained that part of the testing was to stratify the measure results along both individual- and provider-level dimensions by whether the beneficiary is dual-eligible or not.
- The developer explained that the risk-adjusted cost for providers with higher dual-eligible beneficiaries either increases or remains stable. Similarly, the developer noted that the risk-adjusted cost for non-dual episodes increases among providers with higher dual-eligible beneficiaries.
- The developer noted that these results could indicate that providers with higher dual-eligible beneficiaries may perform worse systematically. Therefore, stratification for social risk might be a preferred approach to risk adjustment to avoid unintentionally removing some of the difference in performance due to provider-level effect versus individual-level effect.
- One Standing Committee member questioned why the measure excludes providers with less than 10 episodes per period, as the size of the episode count and the experience of a surgeon could both vary according to social risk factors. The developer responded by stating that they had not performed that specific analysis.
- One Standing Committee member questioned whether the developer considered risk-adjusting by geographical location (i.e., urban versus rural; high resource versus low resource). The developer explained that they looked at provider characteristics at the Taxpayer Identification Number (TIN) and TIN/National Provider Identifier (NPI) level and noted that the results were remarkably similar in urban and rural locations.
- One Standing Committee member expressed concern with the small size of the TEP, as 11 panel members might not be a sufficient representation of the total number of orthopedic surgeons in the U.S.
- The developer responded by detailing an iterative process to measure development that included an initial TEP of a more significant size (n=29 members) and included experts in musculoskeletal disease management with affiliations in 26 organizations and specialty societies to prioritize which measures would be the most impactful for this area of care.
- The developer continued to note that they received input and expertise through a second TEP (n=11), national field-testing, and a commenting period to which they received 67 responses from attributed clinicians.
- Lastly, one Standing Committee member raised a concern: Patients who expire should be included within the measure and not excluded from the measured population, noting that clinicians should be accountable for this significant complication.

3. Feasibility: Total votes-10; H-7; M-3; L-0; I-0

(3a. Clinical data generated during care delivery; 3b. Electronic sources; 3c. Susceptibility to inaccuracies/ unintended consequences identified; 3d. Data collection strategy can be implemented)

Rationale:

- The Standing Committee agreed that the data elements required for the measure are readily available and could be captured without undue burden and passed the measure on feasibility.

4. Usability and Use:

(Used and useful to the intended audiences for 4a. Accountability and Transparency; 4b. Improvement; and 4c. Benefits outweigh evidence of unintended consequences)

4a. Use**: Total votes-10; Pass-9, No Pass-1**; 4b. Usability: **Total votes-10; H-1; M-6; L-3; I-0**

Rationale:

- The Standing Committee acknowledged that this is a new measure and that the developer did not provide any improvement data.
- One Standing Committee member questioned whether physicians could improve on this measure when over 50 percent of clinicians in the U.S. practice in a hospital or healthcare system that often dictates where patients go for diagnostic and preoperative testing and surgical procedures.
- The Standing Committee expressed caution because it is not clear how individual, or clinician-group practices can improve on the measure while increasing or maintaining quality. Specifically, if needed services are withheld or moved to alternative lower-cost care settings, monitoring should be in place to ensure appropriate care is provided.
- The developer reiterated that field-testing reports are provided to all clinicians. In these reports, cost performance is broken into distinct performance categories, such as complications and post-acute care use. Providers are compared to a national average and a set of providers with a similar risk composition.
- A Standing Committee noted that the measure will be used in the CMS MIPS program and questioned how attributed clinicians would differentiate between the different care settings (i.e., ambulatory surgical centers [ACS], inpatient, and office).
- The developer responded by explaining that detailed episodic information, including the standardized cost for all services, is provided at the patient level in MIPS feedback reports.

5. Related and Competing Measures

- No related or competing measures were noted.

6. Standing Committee Recommendation for Endorsement: **Total Votes- 10; Y-8; N- 2**

7. Public and Member Comment

- One public comment not in favor of the measure was submitted prior to the measure evaluation. This public commenter expressed several concerns with the signal-to-noise reliability statistics and low reliability thresholds, the correlation between the cost measures and any one quality measure within the MIPS program, and the risk adjustment methodology.
- No public or NQF member comments were received during the measure evaluation meeting.
- One public comment not in favor of the measure was submitted following the measure evaluation meeting.
    - The commenter noted opposition to NQF #3623 due to the lack of correlations of the cost measure with quality measures and the omission of social risk factors in the risk adjustment model. Specifically, the commenter suggested that testing should demonstrate use in MIPS to make meaningful distinctions in costs associated with the care provided to the patients. Additionally, the commenter noted the testing results (i.e., accountable-entity reliability, empirical validity, and risk adjustment) do not provide the information to ensure the measure provides the desired results, such as:
        - it does not meet the minimum acceptable threshold of 0.7 for the accountable-entity reliability;
        - the empirical validity testing does not include an assessment of this measure with a quality measure;
        - the current risk adjustment model is not adequate due to the adjusted R-squared result of 0.160, nor is the measure adequately tested and adjusted for social risk factors; and
        - the testing provided in Section 2b4. *Identification of Statistically Significant and Meaningful Differences in Performance* does not directly address whether the costs

attributed to physicians and practices enable us to distinguish between low and high performers.

- The developer responded by noting that each of the measures has high reliability at both the TIN [0.86] and TIN-NPI [0.80] levels. The developer continued by highlighting that these [TIN and TIN-NPI] exceed the 0.4 mean reliability standard established through rulemaking for cost measures in MIPS. Furthermore, the developer addressed the commenter's concern (i.e., testing results should demonstrate its reliability for use of the measure in MIPS) by stating that the threshold set by CMS through regulatory processes is pertinent to any evaluation of reliability and noted the SMP passed all measures on the reliability criterion. The developer noted the SMP rated the reliability as high (H: 7, M: 1).
- The Standing Committee did not raise any concerns with the comments, nor did it raise any concerns with the developer's response and maintained its decision to recommend the measure for endorsement.

8. Consensus Standards Approval Committee (CSAC) Endorsement Decision: Total votes- 15; Yes-15; No-0
December 9, 2022: Endorsed

- The CSAC upheld the Standing Committee's decision to recommend the measure for endorsement.

9. Appeals

- No appeals were received.

NQF #3625 Non-Emergent Coronary Artery Bypass Graft (CABG) Measure (CMS/Acumen, LLC)

Measure Worksheet | Specifications

Description: The Non-Emergent CABG episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo a CABG procedure during the performance period. The measure score is the clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from 30 days prior to the clinical event that opens, or "triggers," the episode through 90 days after the trigger. Patient populations eligible for the Non-Emergent CABG measure include Medicare beneficiaries enrolled in Medicare Parts A and B.

Numerator Statement: Not required for cost measures.

Denominator Statement: Not required for cost measures.

Exclusions: Exclusions are used in the Non-Emergent CABG Measure to ensure a homogenous and comparable patient population within the measure's focus on non-emergent CABG procedures. These exclusions focus on removing patients where fair comparisons cannot be made across providers, preventing potential threats to measure validity and ensuring that episodes provide meaningful information to attributed clinicians. These exclusions are listed below:

- Episodes where patient death date occurred before the episode end.
- Episodes where the trigger claim was not in an acute IP hospital setting as defined by subsection (d) or in an acute hospital in Maryland.
- Episodes where the inpatient stay did not have a relevant MS-DRG code.
- Episodes that included an emergent CABG procedure.
- Episodes that included a concurrent cox maze procedure.
- Episodes in which the patient was on dialysis for end-stage renal disease (ESRD).
- Episodes in which the patient was in shock prior to the CABG procedure.
- Episodes that included a redo sternotomy.
- Episodes classified as outlier cases.

Further explanation and rationale for each of the measure exclusions above can be found in Section S.9.1 of the Intent to Submit form. Please also see Section 2b6 (Missing Data Analysis and Minimizing Bias) of this testing form for more information on exclusions implemented as part of data processing.

Given the rationale for the exclusions noted above, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a higher or lower mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). To demonstrate this, we examined the distributions of observed cost and ratio of observed over expected spending (calculated by applying existing risk factor coefficients to the excluded episodes) for each excluded population. We then compared the cost characteristics of the excluded episodes to that of episodes included in the measure to assess the distinctness between the two patient cohorts.

Adjustment/Stratification: Stratification by risk category/subgroup

Level of Analysis: Clinician: Group/Practice, Clinician: Individual

Setting of Care: Inpatient/Hospital

Type of Measure: Cost and Resource use

Data Source: Claims

Measure Steward: CMS

STANDING COMMITTEE MEETING July 12, 2022

1. Importance to Measure and Report:

(1a. High Impact, 1b. Opportunity for Improvement)

1a. High Impact and Opportunity for Improvement: **Total votes- 10; H- 1; M-8; L-1; I- 0**

Rationale:

- The Standing Committee reviewed data demonstrating a high prevalence of nonemergent CABG surgeries among Medicare beneficiaries reflecting substantial Medicare expenditures.
- The Standing Committee noted the downward trajectory and steady decline of CABG cases and mortality. The developer explained that with the advancements in interventional cardiology, the number of CABG procedures would continue to decrease.
- During the discussion on opportunities for improvement, the Standing Committee noted that the performance gap data indicated a mean score of 1.01 (SD of 0.09, IQR of 0.09) at the clinician-group level and a mean score of 1.00 (SD of 0.08, IQR of 0.09) at the individual-clinician level.
- One Standing Committee member noted that the measure aims to reduce the number of avoidable readmissions and the use of appropriate post-acute care and questioned the rationale for making this measure a cost measure instead of a quality measure. The developer explained that the MIPS cost performance category requires measures based on care episode groups. The developer noted that they selected the CABG episode because it is a high-volume, high-cost care area.
- The Standing Committee accepted the developer's rationale and agreed that this measure captures an area of high impact and resource use that warrants a national performance measure.

2. Scientific Acceptability of Measure Properties:

(2a. Reliability - precise specifications, testing; 2b. Validity - testing, threats to validity)

2a. Reliability: **Total votes- 10; H-1; M-8; L-1; I-0**; 2b. Validity: **Total votes- 10; H-0; M-7; L-2; I-1**

Rationale:

- The SMP reviewed this measure and passed it with a rating of moderate on both reliability **(Total votes-8; H-4, M-4, L-0, I-0**) and validity **(Total votes-8; H-0, M-5, L-3, I-0).**
- The Standing Committee agreed with the SMP's evaluation, which stated that the developer's signal-to-noise and split-sample reliability testing were appropriate and that the testing results indicated a moderate measure score reliability.
- While the Standing Committee agreed that the reliability testing was robust, one Standing Committee member requested clarification on why the developer selected the 10-episode case minimum. The developer explained that careful consideration was given to both coverage and reliability when determining the case minimum to ensure that smaller providers with lower case volumes are assessed.
- The Standing Committee noted that the developer conducted both empirical validity testing and a systematic reassessment of face validity through a TEP. Although the Standing Committee agreed that the developer's assessment of face validity was robust, it raised some questions regarding the empirical testing.
- One Standing Committee member noted that 33 percent of CABG procedures are performed among the female population. The developer explained that gender-based anatomical and physiological differences

between men and women (i.e., smaller coronary artery size, heart failure with preserved ejection fraction, and postmenopausal estrogen withdrawal) could contribute to a lower prevalence of procedures among women.

- While the Standing Committee did note that the developer evaluated the empirical validity of this measure by examining its correlation with an NQF-endorsed measure of resource use (NQF #2158 *Medicare Spending per Beneficiary [MSPB] Hospital*), it also expressed interest in understanding the correlation between this measure and a quality measure. The developer noted that, as the discussions had during the review of NQF #3623, they performed a correlation analysis with NQF #3495 *Hospital-Wide 30-Day, All-Cause, Unplanned Readmission Rate [HWR] for the Merit-Based Incentive Payment System [MIPS]-Eligible Clinician Groups* (The Pearson correlations at the TIN and TIN-NPI levels were 0.35 and 0.1, respectively).
- One Standing Committee member noted a concern raised during the SMP's evaluation related to the sizable proportion of episodes excluded from the measure. The developer explained that the exclusion logic is designed to capture only nonemergent CABG procedures.
- The Standing Committee noted that the measure's risk model, which includes 110 risk factors, has an adjusted r-squared of 0.44 and that the results from the developer's stepwise analysis did not support the inclusion of social risk factors.
- Overall, the Standing Committee accepted the developer's rationale, agreed that the validity testing was sufficient, and passed the measure on validity.

3. Feasibility: Total votes-10; H-9; M-1; L-0; I-0

(3a. Clinical data generated during care delivery; 3b. Electronic sources; 3c. Susceptibility to inaccuracies/ unintended consequences identified; 3d. Data collection strategy can be implemented)

Rationale:

- The Standing Committee agreed that the data elements required for the measure are readily available and could be captured without undue burden and passed the measure on feasibility.

4. Usability and Use:

(Used and useful to the intended audiences for 4a. Accountability and Transparency; 4b. Improvement; and 4c. Benefits outweigh evidence of unintended consequences)

4a. Use: **Total votes-10; Pass-9; No Pass-1;** 4b. Usability: **Total votes- 10; H-0; M-7; L-3; I-0**

Rationale:

- The Standing Committee noted that the measure is currently used within the QPP MIPS program.
- While the Standing Committee did acknowledge that this is a new measure and that the developer did not provide any improvement data, it raised concern with how the measure's performance results can be used to further improvement in care. Specifically, the Standing Committee questioned how the developer plans to differentiate between natural variation and areas of actual improvement in care.
- The developer noted that they expect an early reduction in cost to occur and then a gradual flattening out and convergence across providers. The developer plans to address improvement over time during the maintenance process.
- The Standing Committee further noted that opportunities for significant cost savings might be missed when outlier cases are eliminated from the data, as this may be where the actual waste and inefficiencies reside. The developer clarified that 1 percent of episodes at both ends of the distribution are excluded based on the residuals (i.e., the difference between expected and overserved cost) and not just excluding high-cost episodes.
- Although several Standing Committee members continued to have concerns about usability and how performance results can be used to improve care quality, they ultimately passed the measure on use and usability.

5. Related and Competing Measures

- No related or competing measures were noted.

6. Standing Committee Recommendation for Endorsement: Total votes-10; Y-9; N- 1

7. Public and Member Comment

- One public comment not in favor of the measure was submitted prior to the measure evaluation. This public commenter expressed several concerns with the signal-to-noise reliability statistics and low reliability thresholds, the correlation between the cost measures and any one quality measure within the MIPs program, and the risk adjustment methodology.
- No public or NQF member comments were received during the measure evaluation meeting.
- One public comment not in favor of the measure was submitted following the measure evaluation meeting.
    - The commenter noted opposition to NQF #3625 due to the lack of correlations of the cost measure with quality measures and the omission of social risk factors in the risk adjustment model. Specifically, the commenter suggested that testing should demonstrate use in MIPS to make meaningful distinctions in costs associated with the care provided to the patients. Additionally, the commenter noted the testing results (i.e., accountable-entity reliability, empirical validity, and risk adjustment) do not provide the information to ensure the measure provides the desired results, such as:
        - it does not meet the minimum acceptable threshold of 0.7 for the accountable-entity reliability;
        - the empirical validity testing does not include an assessment of this measure with a quality measure;
        - the current risk adjustment model is not adequate due to the adjusted R-squared result of 0.160, nor is the measure adequately tested and adjusted for social risk factors; and
        - the testing provided in Section 2b4. *Identification of Statistically Significant and Meaningful Differences in Performance* does not directly address whether the costs attributed to physicians and practices enable us to distinguish between low and high performers.
    - The developer responded by noting that each of the measures has high reliability at both the TIN [0.84] and TIN-NPI [0.75] levels. The developer continued by highlighting that these [TIN and TIN-NPI] exceed the 0.4 mean reliability standard established through rulemaking for cost measures in MIPS. Furthermore, the developer addressed the commenter's concern (i.e., testing results should demonstrate its reliability for use in MIPS) by stating that the threshold set by CMS through regulatory processes is pertinent to any evaluation of reliability and noted the SMP passed all measures on the reliability criterion. The developer noted that half of the SMP rated the reliability as high, while the other half rated the reliability as moderate (H: 4, M: 4).
    - The Standing Committee did not raise any concerns with the comments, nor did it raise any concerns with the developer's response and maintained its decision to recommend the measure for endorsement.

8. Consensus Standards Approval Committee (CSAC) Endorsement Decision: Total votes- 15; Yes-15; No-0
December 9, 2022: Endorsed

- The CSAC upheld the Standing Committee's decision to recommend the measure for endorsement.

9. Appeals

- No appeals were received.

NQF #3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure (CMS/Acumen, LLC)

Measure Worksheet | Specifications

Description: Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure
Numerator Statement: Not required for cost measures.
Denominator Statement: Not required for cost measures.
Exclusions: Exclusions are used in the Lumbar Spine Fusion Measure to ensure a homogenous and comparable patient population within the measure's focus on surgeries for lumbar spine fusion. These exclusions focus on removing patients where fair comparisons cannot be made across providers, preventing potential threats to

measure validity and ensuring that episodes provide meaningful information to attributed clinicians. These exclusions are listed below:

- Episodes where patient death date occurred before the episode end.
- Episodes where the trigger claim was not in an ambulatory/office-based care setting, IP hospital,
- OP hospital, or ASC based on its place of service.
- Episodes with inpatient procedures, where the inpatient stay did not occur in either an acute hospital as defined by subsection (d) or in an acute hospital in Maryland.
- Episodes with inpatient procedures, where the inpatient stay did not have a relevant MS-DRG code.
- Episodes where the patient had cancer.
- Episodes where the patient had an osteoporotic compression fracture.
- Episodes where the patient had an infection
- Episodes where the patient underwent a redo lumbar fusion.
- Episodes where the patient experienced trauma due to fracture.
- Episodes where the patient had scoliosis and/or kyphosis.
- Episodes where the patient had a spinal fusion within 120 days prior to the episode, with the exception of cervical spinal fusions
- Episodes that included procedures with curvature, malignancy, infections, or extensive fusion
- Episodes classified as outlier cases.

Further explanation and rationale for each of the measure exclusions above can be found in Section S.9.1 of the Intent to Submit form. Please also see Section 2b6 (Missing Data Analysis and Minimizing Bias) of this testing form for more information on exclusions implemented as part of data processing.

Given the rationale for the exclusions noted above, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a higher or lower mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). To demonstrate this, we examined the distributions of observed cost and ratio of observed over expected spending (calculated by applying existing risk factor coefficients to the excluded episodes) for each excluded population. We then compared the cost characteristics of the excluded episodes to that of episodes included in the measure to assess the distinctness between the two patient cohorts.

Adjustment/Stratification: Stratification by risk category/subgroup

Level of Analysis: Clinician: Group/Practice, Clinician: Individual

Setting of Care: Ambulatory Care: Clinic/Urgent Care, Inpatient/Hospital, Other

Type of Measure: Cost and Resource use

Data Source: Claims

Measure Steward: CMS

STANDING COMMITTEE MEETING July 12, 2022

1. Importance to Measure and Report:

(1a. High Impact, 1b. Opportunity for Improvement)

1a. High Impact and Opportunity for Improvement: **Total votes- 10; H-4; M-5; L-1; I-0**

Rationale:

- The Standing Committee reviewed data demonstrating a high prevalence of degenerative lumber conditions affecting more than 6 million Medicare patients and a total admission expenditure for lumbar spine fusion surgeries exceeding $3.6 billion in 2013.
- During the discussion on opportunities for improvement, the Standing Committee noted that the performance gap data indicated a mean score of 1.01 (SD of 0.09; IQR of 0.10) at the clinician-group level and a mean score of 1.00 (SD 0.10; IQR 0.11) at the individual-clinician level.
- One Standing Committee member requested the cost-per-case spread between the 10th and 90th percentiles. The developer reported a mean measure score of $12,784 at the TIN and $13,372 at the TIN-NPI level.

- One Standing Committee member questioned what services tend to drive cost-per-case variability. The developer explained that acute readmissions and post-acute care have the most considerable influence on cost (e.g., the mean observed cost with unplanned readmissions is $53,000 compared to the mean observed cost without unplanned readmission is $37,000).
- The Standing Committee agreed that this measure captures an area of high impact and resource use that warrants a national performance measure and passed the measure on both criteria.

2. Scientific Acceptability of Measure Properties:

(2a. Reliability - precise specifications, testing; 2b. Validity - testing, threats to validity)

2a. Reliability: **Total votes- 10; H-3; M-7; L-0; I-0**; 2b. **Validity: Total votes- 10; H-0; M-9; L-0; I-1**

Rationale:

- The SMP reviewed this measure and passed it with a rating of moderate on both reliability **(Total votes-8; H-4, M-4, L-0, I-0**) and validity **(Total votes-8; H-0, M-6, L-2, I-0).**
- The Standing Committee noted that the developer conducted signal-to-noise and split-sample reliability testing and agreed that both approaches were appropriate; it also agreed that the testing results indicated a moderate measure score reliability.
- The Standing Committee noted that the developer conducted both empirical validity testing and a systematic reassessment of face validity via a TEP. While the Standing Committee agreed that the face validity testing was robust, it raised some questions regarding the empirical testing.
- One Standing Committee questioned how the developer determined measure exclusions for specific episodes (e.g., patients with cancer, patients with an infection, or patients who underwent a redo lumbar fusion) to achieve fair comparisons across providers. The developer explained that they convened clinical expert panels to review and vote on services to include within the measure.
- The Standing Committee noted that the measure's risk adjustment model includes 122 risk factors and that the developer included the social risk factors after the base risk adjustment was conducted (i.e., clinical factors).
- One Standing Committee member requested clarification on how risk adjustment was conducted among the three subgroups. The developer explained that they stratified all episodes into three mutually exclusive subgroups (i.e., one subgroup for the three distinct levels of procedures) and applied the risk adjustment model separately within each of the three subgroups.
- The developer further explained that the three subgroup scores are rolled up at the provider level to calculate the overall measure score.
- One Standing Committee member noted that it is difficult to parse out data within the current risk model related to base and race when the model includes base, dual eligibility status, and race. The Standing Committee member suggested that the developer consider a risk model that only provides base plus race.
- The Standing Committee agreed that the validity testing was appropriate and ultimately passed the measure on both criteria.

3. Feasibility: Total votes- 10; H-8; M-2; L-0; I-0

(3a. Clinical data generated during care delivery; 3b. Electronic sources; 3c. Susceptibility to inaccuracies/ unintended consequences identified; 3d. Data collection strategy can be implemented)

Rationale:

- The Standing Committee agreed that the data elements required for the measure are readily available and could be captured without undue burden and passed the measure on feasibility.

4. Usability and Use:

(Used and useful to the intended audiences for 4a. Accountability and Transparency; 4b. Improvement; and 4c. Benefits outweigh evidence of unintended consequences)

4a. Use: **Total votes- 10; Pass-10, No Pass-0;** 4b. Usability: **Total votes-10; H-1; M-7; L-1; I-1**

Rationale:

- The Standing Committee acknowledged that the measure is currently used within the QPP MIPS program.
- The Standing Committee acknowledged that this is a new measure and that the developer did not provide any improvement data.
- One Standing Committee member raised concern that the developer did not indicate any unintended consequences when racial disparities and undertreatment exist.

- Furthermore, the Standing Committee member noted that there is no explicit tie to current quality.
- The developer explained that the cost drivers are related to adverse outcomes; undertreatment typically results in costly adverse events that the measure will capture within the 90-day postoperative period.
- The Standing Committee member further questioned what events contribute to higher costs (e.g., readmissions, laboratory tests, and diagnostic imaging). The developer noted that while they were unable to share the information during the meeting, the information is available in the field-testing reports provided to clinicians, and they will consider including a summarization of the breakdown during the maintenance review.
- Another Standing Committee member highlighted that the use of medications, particularly opioid pain medications, are not considered pre- and post-procedure for this condition. The developer noted that drugs are included in the service assignment and further highlighted the importance of opioid use quality measures, which look specifically at prescribing practices and use.
- One Standing Committee member questioned why the developer did not include prescription drug data. The developer explained that standardized Medicare Part D drug costs were unavailable at the time of measure development. Furthermore, the developer noted that the workgroups were concerned with the variation in drug prices, which was not within the clinician's purview of control.
- The Standing Committee passed the measure on use and usability.

5. Related and Competing Measures

- No related or competing measures were noted.

6. Standing Committee Recommendation for Endorsement: Total votes- 10; Y-8; N- 2

7. Public and Member Comment

- One public comment not in favor of the measure was submitted prior to the measure evaluation. This public commenter expressed several concerns with the signal-to-noise reliability statistics and low reliability thresholds, the correlation between the cost measures and any one quality measure within the MIPs program, and the risk adjustment methodology.
- No public or NQF member comments were received during the measure evaluation meeting.
- One public comment not in favor of the measure was submitted following the measure evaluation meeting.
    - The commenter noted opposition to NQF #3626 due to the lack of correlations of the cost measure with quality measures and the omission of social risk factors in the risk adjustment model. Specifically, the commenter suggested that testing should demonstrate use in MIPS to make meaningful distinctions in costs associated with the care provided to the patients. Additionally, the commenter noted the testing results (i.e., accountable-entity reliability, empirical validity, and risk adjustment) do not provide the information to ensure the measure provides the desired results, such as:
        - it does not meet the minimum acceptable threshold of 0.7 for the accountable-entity reliability;
        - the empirical validity testing does not include an assessment of this measure with a quality measure;
        - the current risk adjustment model is not adequate due to the adjusted R-squared result of 0.160, nor is the measure adequately tested and adjusted for social risk factors; and
        - the testing provided in Section 2b4. *Identification of Statistically Significant and Meaningful Differences in Performance* does not directly address whether the costs attributed to physicians and practices enable us to distinguish between low and high performers.
    - The developer responded by noting that each of the measures has high reliability at both the TIN [0.78] and TIN-NPI [0.72] levels. The developer continued by highlighting that these [TIN and TIN-NPI] exceed the 0.4 mean reliability standard established through rulemaking for cost measures in MIPS. Furthermore, the developer addressed the commenter's concern (i.e., testing results should demonstrate its reliability for use of the measure in MIPS) by stating that the threshold set by CMS through regulatory processes is pertinent to any evaluation of reliability and noted

the SMP passed all measures on the reliability criterion. The developer noted that half of the SMP rated the reliability as high, while the other half rated the reliability as moderate (H: 4, M: 4).

- The Standing Committee did not raise any concerns with the comments, nor did it raise any concerns with the developer's response and maintained its decision to recommend the measure for endorsement.

8. Consensus Standards Approval Committee (CSAC) Endorsement Decision: Total votes- 15; Yes-15; No-0 December 9, 2022: Endorsed

- The CSAC upheld the Standing Committee's decision to recommend the measure for endorsement.

9. Appeals

- No appeals were received.

Appendix B: Cost and Efficiency Portfolio—Use in Federal Programs*

NQF#	Title	Federal Programs (Finalized or Implemented)
1598	Total Resource Use Population-Based PMPM Index	None
1604	Total Cost of Care Population-Based PMPM Index	None
2158	Medicare Spending per Beneficiary (MSPB)	Care Compare
2431	Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode of Care for Acute Myocardial Infarction (AMI)	Care Compare Hospital Inpatient Quality Reporting
2436	Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode of Care for Heart Failure (HF)	Care Compare Hospital Inpatient Quality Reporting
2579	Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode of Care for Pneumonia	Care Compare Hospital Inpatient Quality Reporting
3474	Hospital-Level, Risk-Standardized Payment Associated With a 90-Day Episode of Care for Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)	None
3509	Routine Cataract Removal With Intraocular Lens (IOL) Implantation	None
3510	Screening/Surveillance Colonoscopy	None
3512	Knee Arthroplasty	None
3561	Medicare Spending per Beneficiary Post-Acute Care Measure for Inpatient Rehabilitation Facilities	Care Compare
3562	Medicare Spending per Beneficiary Post-Acute Care Measure for Long-Term Care Hospitals	Care Compare Long-Term Care Hospital Quality Reporting
3575	Total per Capita Cost (TPCC)	None
3623	Elective Primary Hip Arthroplasty	Quality Payment Program (QPP) Merit-based Incentive Payment System (MIPS)
3625	Non-Emergent Coronary Artery Bypass Graft (CABG)	Quality Payment Program (QPP) Merit-based Incentive Payment System (MIPS)
3626	Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels	Quality Payment Program (QPP) Merit-based Incentive Payment System (MIPS)

*Abstracted from the CMS Measures Inventory Tool Last Accessed on January 17, 2023.

Appendix C: Cost and Efficiency Standing Committee and NQF Staff

STANDING COMMITTEE

Sunny Jhamnani, MD (Co-Chair)
Provider, Dignity Health
Phoenix, Arizona

Kristine Martin Anderson, MBA (Co-Chair)
President, Civilian Sector, Booz Allen Hamilton
Bethesda, Maryland

Robert Bailey, MD
Senior Director, Janssen Scientific Affairs, LLC
Titusville, New Jersey

Bijan Borah, MSc, PhD
Mayo Clinic, College of Medicine
Rochester, Minnesota

Cory Byrd
Humana, Inc.
Louisville, Kentucky

Amy Chin, MS
Assistant Vice President, Hospital for Special Surgery
New York City, New York

Lindsay Erickson, MPH *(Inactive)*
Integrated Healthcare Association (IHA)
Oakland, California

Risha Gidwani, DrPH
Senior Policy Researcher/Adjunct Associate Professor, RAND Corporation/UCLA School of Public Health
Santa Monica, California

Emma Hoo
Pacific Business Group on Health (PBGH)
San Francisco, California

Sean Hopkins, BS
New Jersey Hospital Association
Princeton, New Jersey

Jonathan Jaffrey, MD, MS, MMM

Chief Population Health Officer/President ACO, University of Wisconsin School of Medicine and Public Health Madison, Wisconsin

Dinesh Kalra, MD
Director, Rush University
Chicago, Illinois

Suman Majumdar, PhD (*Inactive*)
Financial Analytics Manager, Washington State Healthcare Authority
Olympia, WA

Alefiyah Mesiwala, MD, MPH *(Inactive)*
Senior Medical Director for Value-based Care and Innovation, UPMC Health Plan
Pittsburgh, Pennsylvania

Pamela Roberts, PhD, OTR/L, SCFES, FAOTA, CPHQ, FNAP, FACRM
Executive Director and Professor Physical Medicine and Rehabilitation, Executive Director to the Office of the Chief Medical Officer, and Co-Director Division of Informatics in the Department of Biomedical Sciences, Cedars-Sinai Medical Center
Los Angeles, California

Mahil Senathirajah, MBA
IBM Watson Health
Santa Barbara, California

Matthew Titmuss, DPT
Assistant Vice President, Hospital for Special Surgery
New York, New York

Danny van Leeuwen, Opa, RN, MPH
Health Hats
Arlington, Virginia

NQF STAFF

Elizabeth Drye, MD, SM
Chief Scientific Officer, Measurement Science and Application

Tricia Elliot, DHA, MBA, CPHQ, FNAHQ
Vice President, Measurement Science and Application (*Former*)

Poonam Bal, MHSA
Senior Director, Measurement Science and Application (*Former*)

Matthew Pickering, PharmD
Managing Director, Measurement Science and Application

Laura Blum Meisnere, MA
Senior Director, Measurement Science and Application

Udara Perera, DrPHc, MPH
Director, Measurement Science and Application

LeeAnn White, MS, BSN
Director, Measurement Science and Application (*Former*)

Isaac Sakyi, MSGH
Manager, Measurement Science and Application

Tristan Wind, BS, ACHE-SA
Analyst, Measurement Science and Application

Karri Albanese, BA
Analyst, Measurement Science and Application *(Former)*

Matilda Epstein, MPH
Associate, Measurement Science and Application

Kate Murphy, BS
Associate, Measurement Science and Application

Victoria Quinones, AA, PMP
Project Manager, Program Operations

Taroon Amin, PhD
Consultant

Appendix D: Measure Specifications

NQF #3623 Elective Primary Hip Arthroplasty Measure Elective Primary Hip Arthroplasty Measure

STEWARD

Centers for Medicare & Medicaid Services

DESCRIPTION

The Elective Primary Hip Arthroplasty episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who receive an elective primary hip arthroplasty during the performance period. The measure score is a clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from the 30 days prior to the clinical event that opens or "triggers" the episode, through 90 days after the trigger. Patient populations eligible for the Elective Primary Hip Arthroplasty measure include Medicare beneficiaries enrolled in Medicare Parts A and B.

TYPE

Cost and Resource use

DATA SOURCE

Claims

LEVEL

Clinician: Group/Practice, Clinician: Individual

SETTING

Inpatient/Hospital, Ambulatory Care: Clinic/Urgent Care, Ambulatory Care: Clinician Office, Other

EXCLUSIONS

Exclusions are used in the Hip Arthroplasty Measure to ensure a homogenous and comparable patient population within the measure's focus on elective primary hip arthroplasties. These exclusions focus on removing patients where fair comparisons cannot be made across providers, preventing potential threats to measure validity and ensuring that episodes provide meaningful information to attributed clinicians. These exclusions are listed below:

- Episodes where patient death date occurred before the episode end.
- Episodes where the trigger claim was not in an ambulatory/office-based care setting, IP hospital, OP hospital, or ASC based on its place of service.
- Episodes with inpatient procedures, where the inpatient stay did not occur in either an acute hospital as defined by subsection (d) or in an acute hospital in Maryland.
- Episodes with inpatient procedures, where the inpatient stay did not have a relevant MS-DRG code.
- Episodes in which the patient underwent a staged or same-day bilateral hip arthroplasty.
- Episodes where the hip replacement was performed due to cancer, hip fracture, or trauma.
- Episodes where the patient had a congenital deformity of the hip, osteomyelitis of the hip or femur, or a septic joint.
- Episodes classified as outlier cases.

Further explanation and rationale for each of the measure exclusions above can be found in Section S.9.1 of the Intent to Submit form. Please also see Section 2b6 (Missing Data Analysis

and Minimizing Bias) of this testing form for more information on exclusions implemented as part of data processing.

Given the rationale for the exclusions noted above, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a higher or lower mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). To demonstrate this, we examined the distributions of observed cost and ratio of observed over expected spending (calculated by applying existing risk factor coefficients to the excluded episodes) for each excluded population. We then compared the cost characteristics of the excluded episodes to that of episodes included in the measure to assess the distinctness between the two patient cohorts.

RISK ADJUSTMENT

Statistical risk model

STRATIFICATION

Differences in case mix are controlled for using an evidence-based statistical risk model with 121 risk factors, including patient health status and clinical factors. This measure's risk adjustment model is not stratified by risk categories.

The risk adjustment model for the Elective Primary Hip Arthroplasty measure broadly follows the CMS-HCC risk adjustment methodology, which is derived from Medicare Parts A and B claims and is used in the Medicare Advantage (MA) program. Although the MA risk adjustment model includes 24 age/sex variables, this risk adjustment model does not adjust for sex and so only includes 12 age categorical variables. Severity of illness is measured using HCCs, indicators of enrollment and long-term care status, and disease interactions. The risk adjustment model also includes variables for additional factors affecting resource use for the measure, identified based on input from the expert clinician workgroup.

The model includes 79 HCC indicators derived from the patient's Parts A and B claims during the period 120 days prior to the episode trigger and are specified in the CMS-HCC V22 2016 model. Episodes for patients without a full 120-day lookback period are excluded from the measure. This 120-day period is used to measure patients' health status and ensures that each patient's claims record contains sufficient fee-for-service data for risk adjustment purposes.

In addition, the risk adjustment model includes status indicator variables for whether the patient qualifies for Medicare through Disability or ESRD. The model also includes an indicator of whether the patient recently required long-term care, defined as 90 days in a long-term care facility without being discharged to community for 14 days. Patients who need to reside in long-term care facilities typically require more intensive care than beneficiaries who live in the community. These enrollment and long-term care status variables are non-diagnostic indicators of severity of illness.

The model also accounts for disease interactions between HCCs and/or enrollment status variables included in the MA model. These interactions are included because certain combinations of comorbidities increase costs more than is predicted by the HCC indicators alone. Furthermore, the risk adjustment model includes measure-specific factors intended to

further isolate cost variation to those costs that attributed clinicians can reasonably influence. These additional variables were informed by clinical rationale and input from the expert clinician workgroup, empirical evidence of explanatory power over cost variation, and are present at the start of care to focus on clinical characteristics that are likely out of the reasonable sphere of influence of the attributed clinician.

As with the CMS-HCC model, the risk adjustment approach for this measure uses an ordinary least squares (OLS) linear regression model. The predicted, or expected, cost is winsorized at 0.5th percentile to make sure episodes with unusually small predicted cost, which would lead to abnormally large O/E ratios, do not dominate certain clinicians' final score. The winsorized expected costs are renormalized to ensure the average expected episode cost is the same before and after winsorizing. Then, extremely low- or high-cost outlier episodes with residuals below the 1st percentile or above the 99th percentile are excluded to reduce the effect of these episodes that deviate the most from their expected values in absolute terms. The expected cost after excluding these outliers is again renormalized to ensure that average expected costs are the same after outlier removal.

The Primary Elective Hip Arthroplasty measure accounts for procedures in the following settings: acute inpatient (IP) hospitals, hospital outpatient departments (HOPD), ambulatory/office-based care centers, and ambulatory surgical centers (ASC). The current trigger code is based on CPT/HCPCS codes and does not require an inpatient stay. However, if an inpatient stay is associated with hip arthroplasty, it is included, and the MS-DRG is risk adjusted for. Specifically, an inpatient episode would be included only when the trigger code appears concurrently with MS-DRG 469 or 470, indicating that the hospital stay was for the hip arthroplasty. As total hip arthroplasties are allowed in an outpatient setting, patients who receive a hip arthroplasty in an inpatient setting are likely more complex, and clinicians taking care of these patients should not be penalized for the necessary precaution of a longer inpatient stay.

TYPE SCORE

Ratio

ALGORITHM

N/A

COPYRIGHT / DISCLAIMER

NQF #3625 Non-Emergent Coronary Artery Bypass Graft (CABG) Measure Non-Emergent Coronary Artery Bypass Graft (CABG) Measure

STEWARD

Centers for Medicare & Medicaid Services

DESCRIPTION

The Non-Emergent CABG episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo a CABG procedure during the performance period. The measure score is the clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are

clinically related to the attributed clinician's role in managing care during each episode from 30 days prior to the clinical event that opens, or "triggers," the episode through 90 days after the trigger. Patient populations eligible for the Non-Emergent CABG measure include Medicare beneficiaries enrolled in Medicare Parts A and B.

TYPE

Cost and Resource use

DATA SOURCE

Claims

LEVEL

Clinician: Individual, Clinician: Group/Practice

SETTING

Inpatient/Hospital

EXCLUSIONS

Exclusions are used in the Non-Emergent CABG Measure to ensure a homogenous and comparable patient population within the measure's focus on non-emergent CABG procedures. These exclusions focus on removing patients where fair comparisons cannot be made across providers, preventing potential threats to measure validity and ensuring that episodes provide meaningful information to attributed clinicians. These exclusions are listed below:

- Episodes where patient death date occurred before the episode end.
- Episodes where the trigger claim was not in an acute IP hospital setting as defined by subsection (d) or in an acute hospital in Maryland.
- Episodes where the inpatient stay did not have a relevant MS-DRG code.
- Episodes that included an emergent CABG procedure.
- Episodes that included a concurrent cox maze procedure.
- Episodes in which the patient was on dialysis for end-stage renal disease (ESRD).
- Episodes in which the patient was in shock prior to the CABG procedure.
- Episodes that included a redo sternotomy.
- Episodes classified as outlier cases.

Further explanation and rationale for each of the measure exclusions above can be found in Section S.9.1 of the Intent to Submit form. Please also see Section 2b6 (Missing Data Analysis and Minimizing Bias) of this testing form for more information on exclusions implemented as part of data processing.

Given the rationale for the exclusions noted above, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a higher or lower mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). To demonstrate this, we examined the distributions of observed cost and ratio of observed over expected spending (calculated by applying existing risk factor coefficients to the excluded episodes) for each excluded population. We then compared the cost characteristics of the excluded episodes to that of episodes included in the measure to assess the distinctness between the two patient cohorts.

RISK ADJUSTMENT

Stratification by risk category/subgroup

STRATIFICATION

Differences in case mix are controlled for using an evidence-based statistical risk model with 119 risk factors, including patient health status and clinical factors. The Non-Emergent CABG measure is stratified into two sub-groups:

- CABG with Concurrent Aortic Valve Replacement
- Isolated CABG

By running the risk adjustment model, described below and in Section S.7.2., separately for episodes within each sub-group, the measure accounts for differences in resource use stemming from the type of procedure. This helps ensure that the cost measure is fairly comparing clinicians for CABG overall while preserving clinically meaningful distinctions between the procedure types.

The risk adjustment model for the Non-Emergent CABG measure broadly follows the CMS-HCC risk adjustment methodology, which is derived from Medicare Parts A and B claims and is used in the Medicare Advantage (MA) program. Although the MA risk adjustment model includes 24 age/sex variables, this risk adjustment model does not adjust for sex and so only includes 12 age categorical variables. Severity of illness is measured using HCCs, indicators of enrollment and long-term care status, and disease interactions. The risk adjustment model also includes variables for factors identified by the expert clinician workgroup as affecting resource use.

The model includes 79 HCC indicators derived from the patient's Parts A and B claims during the period 120 days prior to the episode trigger and are specified in the CMS-HCC V22 2016 model. Episodes for patients without a full 120-day lookback period are excluded from the measure. This 120-day period is used to measure patients' health status and ensures that each patient's claims record contains sufficient fee-for-service data for risk adjustment purposes.

In addition, the risk adjustment model includes status indicator variables for whether the patient qualifies for Medicare through Disability or ESRD. The model also includes an indicator of whether the patient recently required long-term care, defined as 90 days in a long-term care facility without being discharged to community for 14 days. Patients who need to reside in long-term care facilities typically require more intensive care than beneficiaries who live in the community. These enrollment and long-term care status variables are non-diagnostic indicators of severity of illness.

The model also accounts for disease interactions between HCCs and/or enrollment status variables included in the MA model. These interactions are included because certain combinations of comorbidities increase costs more than is predicted by the HCC indicators alone. Furthermore, the risk adjustment model includes measure-specific factors intended to further isolate cost variation to those costs that attributed clinicians can reasonably influence. These additional variables were informed by clinical rationale and input from the expert clinician workgroup, empirical evidence of explanatory power over cost variation, and are present at the start of care to focus on clinical characteristics that are likely out of the reasonable sphere of influence of the attributed clinician.

As with the CMS-HCC model, the risk adjustment approach for this measure uses an ordinary least squares (OLS) linear regression model. The predicted, or expected, cost is winsorized at

0.5th percentile to make sure episodes with unusually small predicted cost, which would lead to abnormally large O/E ratios, do not dominate certain clinicians' final score. The winsorized expected costs are renormalized to ensure the average expected episode cost is the same before and after winsorizing. Then, extremely low- or high-cost outlier episodes with residuals below the 1st percentile or above the 99th percentile are excluded to reduce the effect of these episodes that deviate the most from their expected values in absolute terms. The expected cost after excluding these outliers is again renormalized to ensure that average expected costs are the same after outlier removal.

TYPE SCORE

Ratio

ALGORITHM

N/A

COPYRIGHT / DISCLAIMER

NQF #3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure

STEWARD

Centers for Medicare & Medicaid Services

DESCRIPTION

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo surgery for lumbar spine fusion during the performance period. The measure score is the clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from 30 days prior to the clinical event that opens, or "triggers," the episode through 90 days after the trigger. Patient populations eligible for Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure include Medicare beneficiaries enrolled in Medicare Parts A and B.

TYPE

Cost and Resource use

DATA SOURCE

Claims

LEVEL

Clinician: Group/Practice, Clinician: Individual

SETTING

Ambulatory Care: Clinic/Urgent Care, Inpatient/Hospital, Other

EXCLUSIONS

Exclusions are used in the Lumbar Spine Fusion Measure to ensure a homogenous and comparable patient population within the measure's focus on surgeries for lumbar spine fusion. These exclusions focus on removing patients where fair comparisons cannot be made across providers, preventing potential threats to measure validity and ensuring that episodes provide meaningful information to attributed clinicians. These exclusions are listed below:

- Episodes where patient death date occurred before the episode end.
- Episodes where the trigger claim was not in an ambulatory/office-based care setting, IP hospital, OP hospital, or ASC based on its place of service.
- Episodes with inpatient procedures, where the inpatient stay did not occur in either an acute hospital as defined by subsection (d) or in an acute hospital in Maryland.
- Episodes with inpatient procedures, where the inpatient stay did not have a relevant MS-DRG code.
- Episodes where the patient had cancer.
- Episodes where the patient had an osteoporotic compression fracture.
- Episodes where the patient had an infection
- Episodes where the patient underwent a redo lumbar fusion.
- Episodes where the patient experienced trauma due to fracture.
- Episodes where the patient had scoliosis and/or kyphosis.
- Episodes where the patient had a spinal fusion within 120 days prior to the episode, with the exception of cervical spinal fusions
- Episodes that included procedures with curvature, malignancy, infections, or extensive fusion
- Episodes classified as outlier cases.

Further explanation and rationale for each of the measure exclusions above can be found in Section S.9.1 of the Intent to Submit form. Please also see Section 2b6 (Missing Data Analysis and Minimizing Bias) of this testing form for more information on exclusions implemented as part of data processing.

Given the rationale for the exclusions noted above, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a higher or lower mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). To demonstrate this, we examined the distributions of observed cost and ratio of observed over expected spending (calculated by applying existing risk factor coefficients to the excluded episodes) for each excluded population. We then compared the cost characteristics of the excluded episodes to that of episodes included in the measure to assess the distinctness between the two patient cohorts.

RISK ADJUSTMENT

Stratification by risk category/subgroup

STRATIFICATION

Differences in case mix are controlled for using an evidence-based statistical risk model with 122 risk factors, including both patient health status and clinical factors. The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure is stratified into three sub-groups, or mutually exclusive and exhaustive divisions of the overall episode group:

- One-level lumbar fusion
- Two-level lumbar fusion
- Three-level lumbar fusion

By running the risk adjustment model, described below and in Section S.7.2, separately for episodes within each sub-group, the measure accounts for differences in resource use stemming from the complexity of the procedure. This helps ensure that the cost measure is fairly

comparing clinicians for lumbar spine fusion overall while preserving clinically meaningful distinctions within each level.

The risk adjustment model for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure broadly follows the CMS-HCC risk adjustment methodology, which is derived from Medicare Parts A and B claims and is used in the Medicare Advantage (MA) program. Although the MA risk adjustment model includes 24 age/sex variables, this risk adjustment model does not adjust for sex and so only includes 12 age categorical variables. Severity of illness is measured using HCCs, indicators of enrollment and long-term care status, and disease interactions. The risk adjustment model also includes variables for factors identified by the expert clinician workgroup as affecting resource use.

The model includes 79 HCC indicators derived from the patient's Parts A and B claims during the period 120 days prior to the episode trigger and are specified in the CMS-HCC V22 2016 model. Episodes for patients without a full 120-day lookback period are excluded from the measure. This 120-day period is used to measure patients' health status and ensures that each patient's claims record contains sufficient fee-for-service data for risk adjustment purposes.

In addition, the risk adjustment model includes status indicator variables for whether the patient qualifies for Medicare through Disability or ESRD. The model also includes an indicator of whether the patient recently required long-term care, defined as 90 days in a long-term care facility without being discharged to community for 14 days. Patients who need to reside in long-term care facilities typically require more intensive care than beneficiaries who live in the community. These enrollment and long-term care status variables are non-diagnostic indicators of severity of illness.

The model also accounts for disease interactions between HCCs and/or enrollment status variables included in the MA model. These interactions are included because certain combinations of comorbidities increase costs more than is predicted by the HCC indicators alone. Furthermore, the risk adjustment model includes measure-specific factors intended to further isolate cost variation to those costs that attributed clinicians can reasonably influence. These additional variables were informed by clinical rationale and input from the expert clinician workgroup, empirical evidence of explanatory power over cost variation, and are present at the start of care to focus on clinical characteristics that are likely out of the reasonable sphere of influence of the attributed clinician.

As with the CMS-HCC model, the risk adjustment approach for this measure uses an ordinary least squares (OLS) linear regression model. The predicted, or expected, cost is winsorized at 0.5th percentile to make sure episodes with unusually small predicted cost, which would lead to abnormally large O/E ratios, do not dominate certain clinicians' final score. The winsorized expected costs are renormalized to ensure the average expected episode cost is the same before and after winsorizing. Then, extremely low- or high-cost outlier episodes with residuals below the 1st percentile or above the 99th percentile are excluded to reduce the effect of these episodes that deviate the most from their expected values in absolute terms. The expected cost after excluding these outliers is again renormalized to ensure that average expected costs are the same after outlier removal.

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure accounts for procedures in the following settings: acute inpatient (IP) hospitals, hospital outpatient departments (HOPD), ambulatory/office-based care centers, and ambulatory surgical centers (ASC). The current trigger code is based on CPT/HCPCS codes and does not require an inpatient stay. However, risk adjustment for the MS-DRG of the inpatient stay is included, if one is associated with the lumbar spine fusion. Specifically, an inpatient episode would be included only when the trigger code appears concurrently with MS-DRGs 453-455, 459, or 460, indicating that the hospital stay was for the lumbar spine fusion procedure. Furthermore, the measure includes risk adjustment variables for the place of service to account for the significant cost variation across the settings, acknowledging that clinicians may have limited access to different places of service.

TYPE SCORE

Ratio

ALGORITHM

N/A

COPYRIGHT / DISCLAIMER

Appendix E: Related and Competing Measures

There are no related or competing measures.

Appendix F: Pre-Evaluation Comments

Comments received as of June 15, 2022.

NQF #3623 Elective Primary Hip Arthroplasty

Commenter

Koryn Rubin, on behalf of American Medical Association

Comment

The American Medical Association (AMA) appreciates the opportunity to comment on this measure and requests that the Standing Committee carefully consider our comments on its scientific acceptability during this evaluation. The Centers for Medicare and Medicaid Services (CMS) developed this measure specifically for use in the Merit-based Incentive Payment System (MIPS) and we believe that the information and testing provided should demonstrate that its use in MIPS will yield reliable and valid results and enable end users to make meaningful distinctions in the costs associated with the care provided to these patients. The AMA is concerned that the testing results provided, particularly for accountable-entity reliability, empirical validity and the risk adjustment approach, do not provide the information needed to ensure that this measure produces the desired results. Regarding the accountable-entity reliability, we are concerned with the lack of information on reliability results below the 10th percentile, particularly since the scores at the practice and physician levels provided were 0.68 and 0.70 respectively. The AMA believes that the minimum acceptable thresholds should be 0.7 and the measure as specified does not meet this goal. The AMA strongly supports the tenet that cost must be assessed within the context of the quality of care provided; yet, the developer did not demonstrate that this measure correlates to any one quality measure within the MIPS program. We are very troubled that the testing did not include an assessment of this measure with a measure such as the claims-based Risk-Standardized Complication Rate Following Elective Primary Total Hip Arthroplasty and/or Total Knee Arthroplasty (TKA). While we acknowledge that a comparison to this or a similar quality measure will include a broader population, it will provide more meaningful information regarding the validity of the cost measure rather than the current comparison to the Medicare Spending Per Beneficiary measure. Regardless, the AMA does not believe that cost measures against which no quality measure can be assessed should achieve endorsement. The AMA does not believe that the current risk adjustment model is adequate due to the adjusted R-squared result of 0.160 nor is the measure adequately tested and adjusted for social risk factors. It is unclear to us why the developer would test social risk factors after adjusting for clinical risk factors rather than assessing the impact of both clinical and social risk factors in the model at the same time. These variations in how risk adjustment factors are examined could also impact how each variable (clinical or social) perform in the model and remain unanswered questions. In addition, the AMA questions whether the information provided in Section 2b4. Identification of Statistically Significant and Meaningful Differences in Performance is truly useful for accountability and informing patients of the cost of care provided by physicians and practices. Specifically, that the testing does not directly address whether the costs attributed to physicians and practices enable us to distinguish low versus high performers. Since this measure was specifically developed for use in MIPS, analyses of the performance scores using the finalized benchmarking methodology across 10 deciles would provide valuable information on whether the differences in costs between physicians and practices could be considered useful and meaningful. The AMA requests that

these gaps in testing be addressed prior to endorsement of this measure. We appreciate the Committee's consideration of our comments.

NQF #3625 Non-Emergent Coronary Artery Bypass Graft (CABG)

Commenter

Koryn Rubin, on behalf of American Medical Association

Comment

The American Medical Association (AMA) appreciates the opportunity to comment on this measure and requests that the Standing Committee carefully consider our comments on its scientific acceptability during this evaluation. The Centers for Medicare and Medicaid Services (CMS) developed this measure specifically for use in the Merit-based Incentive Payment System (MIPS) and we believe that the information and testing provided should demonstrate that its use in MIPS will yield reliable and valid results and enable end users to make meaningful distinctions in the costs associated with the care provided to these patients. The AMA is concerned that the testing results provided, particularly for accountable-entity reliability, empirical validity and the risk adjustment approach, do not provide the information needed to ensure that this measure produces the desired results. Regarding the accountable-entity reliability, we are concerned with the lack of information on reliability results below the 10th percentile, particularly since the scores at the practice and physician levels provided were 0.69 and 0.64 respectively. The AMA believes that the minimum acceptable thresholds should be 0.7 and the measure as specified does not meet this goal. The AMA strongly supports the tenet that cost must be assessed within the context of the quality of care provided; yet the developer did not demonstrate that this measure correlates to any one quality measure within the MIPS program. We are very troubled that the testing did not include an assessment of this measure with a related quality measure used in MIPS as it would provide more meaningful information regarding the validity of the cost measure rather than the current comparison to the Medicare Spending Per Beneficiary measure. Regardless, the AMA does not believe that cost measures against which no quality measure can be assessed should achieve endorsement. The AMA does not believe that the current risk adjustment model is adequately tested and adjusted for social risk factors. It is unclear to us why the developer would test social risk factors after adjusting for clinical risk factors rather than assessing the impact of both clinical and social risk factors in the model at the same time. These variations in how risk adjustment factors are examined could also impact how each variable (clinical or social) perform in the model and remain unanswered questions. In addition, the AMA questions whether the information provided in Section 2b4. Identification of Statistically Significant and Meaningful Differences in Performance is truly useful for accountability and informing patients of the cost of care provided by physicians and practices. Specifically, that the testing does not directly address whether the costs attributed to physicians and practices enable us to distinguish low versus high performers. Since this measure was specifically developed for use in MIPS, analyses of the performance scores using the finalized benchmarking methodology across 10 deciles would provide valuable information on whether the differences in costs between physicians and practices could be considered useful and meaningful. The AMA requests that these gaps in testing be addressed prior to endorsement of this measure. We appreciate the Committee's consideration of our comments.

NQF #3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels

Commenter

Koryn Rubin, on behalf of American Medical Association

Comment

The American Medical Association (AMA) appreciates the opportunity to comment on this measure and requests that the Standing Committee carefully consider our comments on its scientific acceptability during this evaluation. The Centers for Medicare and Medicaid Services (CMS) developed this measure specifically for use in the Merit-based Incentive Payment System (MIPS) and we believe that the information and testing provided should demonstrate that its use in MIPS will yield reliable and valid results and enable end users to make meaningful distinctions in the costs associated with the care provided to these patients. The AMA is concerned that the testing results provided, particularly for accountable-entity reliability, empirical validity and the risk adjustment approach, do not provide the information needed to ensure that this measure produces the desired results. Regarding the accountable-entity reliability, we are concerned with the lack of information on reliability results below the 10th percentile, particularly since the scores at the practice and physician levels provided were 0.64 and 0.60 respectively. The AMA believes that the minimum acceptable thresholds should be 0.7 and the measure as specified does not meet this goal. The AMA strongly supports the tenet that cost must be assessed within the context of the quality of care provided; yet the developer did not demonstrate that this measure correlates to any one quality measure within the MIPS program. We are very troubled that the testing did not include an assessment of this measure with a related quality measure used in MIPS as it would provide more meaningful information regarding the validity of the cost measure rather than the current comparison to the Medicare Spending Per Beneficiary measure. Regardless, the AMA does not believe that cost measures against which no quality measure can be assessed should achieve endorsement. The AMA does not believe that the current risk adjustment model is adequately tested and adjusted for social risk factors. It is unclear to us why the developer would test social risk factors after adjusting for clinical risk factors rather than assessing the impact of both clinical and social risk factors in the model at the same time. These variations in how risk adjustment factors are examined could also impact how each variable (clinical or social) perform in the model and remain unanswered questions. In addition, the AMA questions whether the information provided in Section 2b4. Identification of Statistically Significant and Meaningful Differences in Performance is truly useful for accountability and informing patients of the cost of care provided by physicians and practices. Specifically, that the testing does not directly address whether the costs attributed to physicians and practices enable us to distinguish low versus high performers. Since this measure was specifically developed for use in MIPS, analyses of the performance scores using the finalized benchmarking methodology across 10 deciles would provide valuable information on whether the differences in costs between physicians and practices could be considered useful and meaningful. The AMA requests that these gaps in testing be addressed prior to endorsement of this measure. We appreciate the Committee's consideration of our comments.

Appendix G: Post-Evaluation Comments

NQF #3623 Elective Primary Hip Arthroplasty Measure (Recommended)

Ms. Koryn Y. Rubin, MHA, American Medical Association

Comment ID#: 8292 (Submitted: 09/26/2022)

Council / Public: HPR

Level of Support: Member Does NOT Support

Comment

The American Medical Association (AMA) agrees with the concerns the Standing Committee expressed regarding the lack of correlations of the cost measures with quality measures as well as the omission of social risk factors in the risk adjustment model. While we are in agreement with these concerns, they are not new and are frequently discussed by this Committee. To repeatedly raise the same concerns with no resolution does not advance our shared goal of representing costs, and ultimately value, and they must be addressed prior to any endorsement of new cost measures. The AMA continues to have concerns with this measure and does not support its endorsement. Specifically, we believe that the following issues must be addressed: • Because this measure was developed for use in the Merit-based Incentive Payment System (MIPS) and we believe that the information and testing provided should demonstrate that its use in MIPS will yield reliable and valid results and enable end users to make meaningful distinctions in the costs associated with the care provided to these patients. • The testing results provided, particularly for accountable-entity reliability, empirical validity, and the risk adjustment approach, do not provide the information needed to ensure that this measure produces the desired results: o It does not meet the minimum acceptable threshold of 0.7 for the accountable-entity reliability; o The empirical validity testing does not include an assessment of this measure with a quality measure; o The current risk adjustment model is not adequate due to the adjusted R-squared result of 0.160 nor is the measure adequately tested and adjusted for social risk factors; and o The testing provided in Section 2b4. Identification of Statistically Significant and Meaningful Differences in Performance does not directly address whether the costs attributed to physicians and practices enable us to distinguish low versus high performers.

Developer Response

Reliability Testing We would like to clarify that each of the measures has high reliability at both the TIN and TIN-NPI levels where the mean in fact exceeds 0.7, as shown in the testing materials. Specifically, the mean reliability for this measure for TIN is 0.86 and for TIN-NPIs is 0.80. This far exceeds the 0.4 mean reliability standard established through rulemaking for cost measures in MIPS. As noted by the commenter, testing results should demonstrate its reliability for use in MIPS; as such, the threshold set by CMS through regulatory processes is pertinent to any evaluation of reliability. Further, we note that the Scientific Methods Panel, whose role is to provide consistency and expertise in the scientific acceptability of measures, passed all measures on the reliability criterion. In fact, the SMP rated the reliability as high (H: 7, M: 1). We reiterate that NQF does not set reliability thresholds, as stated in some of their materials, nor is there agreement in the

literature on a threshold. Please see our submission materials and the CY 2022 PFS final rule (86 FR 65453 – 65454) for further details. Validity Results: Correlations with Related Quality Measures To clarify the commenter's concern about the lack of correlation analyses with quality measures, the empirical validity testing discussed during the Standing Committee evaluation meeting actually did include correlation analyses with quality measures. To recap that discussion for the commenter, we calculated the correlation between the cost measure and a MIPS quality measure for complications after THA and TKA that we constructed using the public specifications. The results confirmed the expected relationship, namely that clinicians who have lower costs tend to have lower rates of complications as demonstrated by medium Pearson correlation of 0.27 at both the TIN and TIN-NPI levels. SRF Testing Methodology To address the comment about the adequacy of SRF testing, we recap the discussion of testing during the Standing Committee evaluation meeting which included additional analyses which reflect guidance from organizations including NQF and ASPE about what considerations should be taken into account when assessing whether or not SRFs should be adjusted. We found that there is little impact on provider scores by risk adjusting for beneficiary dual status. We were however concerned that adjusting for dual status for this measure could risk masking providers' poor performance and exacerbate disparities in care because testing showed that providers who perform worse on dual beneficiaries perform worse on both dual and non-dual patients. That is, provider characteristics are more influencing the higher costs of episodes for patients with dual status, rather than patient factors. The testing approach that we discussed with the Standing Committee is one that has led to the decision to adjust for SRFs when results indicate that it is appropriate to do so. For example, the following two chronic condition measures that were finalized for MIPS 2022 do adjust for dual status: Diabetes and the Asthma/Chronic Obstructive Pulmonary Disease (COPD) episode-based cost measures. Finally, we agree with NQF's comment in the Draft Report that measures must be reviewed on a case-by-case basis to understand whether adjusting for SRFs is appropriate, to avoid unintended consequences. Low R-Squared An R-squared may be low because observed cost is due to provider choice, not beneficiary characteristics. This can point to the need for a cost measure. R-squared metrics must be interpreted within the context of the measure construction, what it is intended to capture, and its use. For example, the measure does not include dialysis services because they are outside of the reasonable influence of the surgeon performing this procedure. If the measure did include dialysis - a costly service - then more variation in observed cost due to dialysis would be explained by the ESRD risk adjustor, yet would not make the measure more "valid". Attributed orthopedic/cardiothoracic/neuro surgeons may in fact consider it to be less "valid" to be held accountable for the costs of dialysis. As such, a low R-squared is conceptually neither required nor expected for a "valid" measure, so some valid measures will have low R-squareds, while others will have high R-squareds. We also note that extensive testing demonstrates the validity of the risk adjustment models for the measure, with model discrimination and calibration results demonstrating good predictive ability across the full range of episodes, from low to high spending risk (Sections 2b3.7-10). There was no evidence of excessive under- or over-estimation at the extremes of episode risk. Information in the cost measure meaningfully distinguishes between performance To confirm, the purpose of section 2b4 of the testing form is to demonstrate that there is clinically and practically significant variation in the measure scores. Given that testing results do show that this variation is present for the measure, they suggest that there are differences in performance, where some clinicians have low performance on the measure and

some clinicians have high performance on the measure. We refer the commenter to other sections of the testing form to address the question of whether the costs included in the measure can meaningfully distinguish between high and low performance. Section 2b1 of the testing form describes how we convened a group of experts to provide detailed input on the measure specifications, including determining clinically related services that should be assigned to the measure. To gather a formal record of the workgroup's systematic input throughout development, workgroup members completed a face validity survey to assess the measure's ability to fulfill its intent to meaningfully compare and evaluate clinicians on cost efficiency. The results of the face validity vote showed that there was overall consensus agreement that the measure can distinguish good from poor performance. Finally, we share the commenter's interest in ensuring that end users can use the information from cost measures. Currently, MIPS participants receive patient-level episode-based cost measure reports which include the following information: episode identifiers (e.g., trigger date); list of all services rendered during the episode and the standardized costs, organized into service categories (e.g., post-trigger costs for outpatient facility costs); patient information (e.g., HCC risk score, sex). CMS will continue to consider feedback about what information is most useful for clinicians.

NQF Response

N/A

NQF Committee Response

Thank you for your comments. The Standing Committee recognizes that cost and resource use measures should be used in the context of and reported with quality measures. The Standing Committee discussed the relationship between cost and quality measures, emphasizing the importance of reporting performance to demonstrate improvements in cost while ensuring similar or higher levels of care quality. However, NQF criteria do not currently require that a cost measure be correlated with a quality measure. Rather, empirical validity testing should demonstrate that the measure's data elements are correct and/or the measure score correctly reflects the cost of care or resources provided. The Standing Committee further notes the need to ensure that providers serving people with SRFs are not penalized unfairly due to a lack of social risk adjustment. While the developer tested for social risk factors (SRFs) for the measure's risk adjustment model, some of the measures under review did not include these SRFs in the final model. Although the Standing Committee recognizes the importance of maximizing the predictive value of a risk adjustment model, elements of a risk model should be included or excluded based on a conceptual and empirical rationale. Thus, the Standing Committee considered the developer's empirical reliability and validity testing, including the Scientific Methods Panel's (SMP) input on both the reliability and validity testing, and the approach to the risk adjustment modeling and agreed to recommend these measures for endorsement.

NQF #3625 Non-Emergent Coronary Artery Bypass Graft (CABG) Measure (Recommended)

Ms. Koryn Y. Rubin, MHA, American Medical Association

Comment ID#: 8293 (Submitted: 09/26/2022)

Council / Public: HPR

Level of Support: Member Does NOT Support

Comment

The American Medical Association (AMA) agrees with the concerns the Standing Committee expressed regarding the lack of correlations of the cost measures with quality measures as well as the omission of social risk factors in the risk adjustment model. While we are in agreement with these concerns, they are not new and are frequently discussed by this Committee. To repeatedly raise the same concerns with no resolution does not advance our shared goal of representing costs, and ultimately value, and they must be addressed prior to any endorsement of new cost measures. The AMA continues to have concerns with this measure and does not support its endorsement. Specifically, we believe that the following issues must be addressed: • Because this measure was developed for use in the Merit-based Incentive Payment System (MIPS) and we believe that the information and testing provided should demonstrate that its use in MIPS will yield reliable and valid results and enable end users to make meaningful distinctions in the costs associated with the care provided to these patients. • The testing results provided, particularly for accountable-entity reliability, empirical validity, and the risk adjustment approach, do not provide the information needed to ensure that this measure produces the desired results: o It does not meet the minimum acceptable threshold of 0.7 for the accountable-entity reliability; o The empirical validity testing does not include an assessment of this measure with a quality measure; o The current risk adjustment model does not adequately test and adjust for social risk factors; and o The testing provided in Section 2b4. Identification of Statistically Significant and Meaningful Differences in Performance does not directly address whether the costs attributed to physicians and practices enable us to distinguish low versus high performers.

Developer Response

Reliability Testing We would like to clarify that each of the measures has high reliability at both the TIN and TIN-NPI levels where the mean in fact exceeds 0.7, as shown in the testing materials. Specifically, the mean reliability for this measure for TINs is 0.84 and for TIN-NPIs is 0.75. This far exceeds the 0.4 mean reliability standard established through rulemaking for cost measures in MIPS. As noted by the commenter, testing results should demonstrate its reliability for use in MIPS; as such, the threshold set by CMS through regulatory processes is pertinent to any evaluation of reliability. Further, we note that the Scientific Methods Panel, whose role is to provide consistency and expertise in the scientific acceptability of measures, passed all measures on the reliability criterion. In fact, half of the SMP members rated the reliability as high, while the other half rated it as moderate (H: 4, M: 4). We reiterate that NQF does not set reliability thresholds, as stated in some of their materials, nor is there agreement in the literature on a threshold. Please see our submission materials and the CY 2022 PFS final rule (86 FR 65453 – 65454) for further details. Validity Results: Correlations with Related Quality Measures To clarify the commenter's concern about the lack of correlation analyses with quality measures, the empirical validity testing discussed during the Standing Committee evaluation meeting actually did include correlation analyses with quality measures. To recap that discussion for the commenter, we calculated the correlation between the cost measure and a MIPS quality measure for unplanned readmissions that we constructed using the public specifications. The results confirmed the expected

relationship, namely that clinicians who have lower costs tend to have lower rates of unplanned readmissions, as demonstrated by the medium to high Pearson correlation between the cost measure and the unplanned readmissions quality measure: 0.35 correlation at the TIN level and 0.41 at the TIN-NPI level. SRF Testing Methodology To address the comment about the adequacy of SRF testing, we recap the discussion of testing during the Standing Committee evaluation meeting which included additional analyses which reflects guidance from organizations including NQF and ASPE about what considerations should be taken into account when assessing whether or not SRFs should be adjusted. We found that there is little impact on provider scores by risk adjusting for beneficiary dual status. We were however concerned that adjusting for dual status for this measure could risk masking providers' poor performance and exacerbate disparities in care because testing showed that providers who perform worse on dual beneficiaries perform worse on both dual and non-dual patients. That is, provider characteristics are more influencing the higher costs of episodes for patients with dual status, rather than patient factors. The testing approach that we discussed with the Standing Committee is one that has led to the decision to adjust for SRFs when results indicate that it is appropriate to do so. For example, the following two chronic condition measures that were finalized for MIPS 2022 do adjust for dual status: Diabetes and the Asthma/Chronic Obstructive Pulmonary Disease (COPD) episode-based cost measures. Finally, we agree with NQF's comments in the Draft Report that measures must be reviewed on a case-by-case basis to understand whether adjusting for SRFs is appropriate, to avoid unintended consequences. Information in the cost measure meaningfully distinguishes between performance To confirm, the purpose of section 2b4 of the testing form is to demonstrate that there is clinically and practically significant variation in the measure scores. Given that testing results do show that this variation is present for the measure, they suggest that there are differences in performance, where some clinicians have low performance on the measure and some clinicians have high performance on the measure. We refer the commenter to other sections of the testing form to address the question of whether the costs included in the measure can meaningfully distinguish between high and low performance. Section 2b1 of the testing form describes how we convened a group of experts to provide detailed input on the measure specifications, including determining clinically related services that should be assigned to the measure. To gather a formal record of the workgroup's systematic input throughout development, workgroup members completed a face validity survey to assess the measure's ability to fulfill its intent to meaningfully compare and evaluate clinicians on cost efficiency. The results of the face validity vote showed that there was overall consensus agreement that the measure can distinguish good from poor performance. Finally, we share the commenter's interest in ensuring that end users can use the information from cost measures. Currently, MIPS participants receive patient-level episode-based cost measure reports which include the following information: episode identifiers (e.g., trigger date); list of all services rendered during the episode and the standardized costs, organized into service categories (e.g., post-trigger costs for outpatient facility costs); patient information (e.g., HCC risk score, sex). CMS will continue to consider feedback about what information is most useful for clinicians.

NQF Response

N/A

NQF Committee Response

Thank you for your comments. The Standing Committee recognizes that cost and resource use measures should be used in the context of and reported with quality measures. The Standing Committee discussed the relationship between cost and quality measures, emphasizing the importance of reporting performance to demonstrate improvements in cost while ensuring similar or higher levels of care quality. However, NQF criteria do not currently require that a cost measure be correlated with a quality measure. Rather, empirical validity testing should demonstrate that the measure's data elements are correct and/or the measure score correctly reflects the cost of care or resources provided. The Standing Committee further notes the need to ensure that providers serving people with SRFs are not penalized unfairly due to a lack of social risk adjustment. While the developer tested for social risk factors (SRFs) for the measure's risk adjustment model, some of the measures under review did not include these SRFs in the final model. Although the Standing Committee recognizes the importance of maximizing the predictive value of a risk adjustment model, elements of a risk model should be included or excluded based on a conceptual and empirical rationale. Thus, the Standing Committee considered the developer's empirical reliability and validity testing, including the Scientific Methods Panel's (SMP) input on both the reliability and validity testing, and the approach to the risk adjustment modeling and agreed to recommend these measures for endorsement.

NQF #3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure (Recommended)

Ms. Koryn Y. Rubin, MHA, American Medical Association

Comment ID#: 8294 (Submitted: 09/26/2022)

Council / Public: HPR

Level of Support: Member Does NOT Support

Comment

The American Medical Association (AMA) agrees with the concerns the Standing Committee expressed regarding the lack of correlations of the cost measures with quality measures as well as the omission of social risk factors in the risk adjustment model. While we are in agreement with these concerns, they are not new and are frequently discussed by this Committee. To repeatedly raise the same concerns with no resolution does not advance our shared goal of representing costs, and ultimately value, and they must be addressed prior to any endorsement of new cost measures. The AMA continues to have concerns with this measure and does not support its endorsement. Specifically, we believe that the following issues must be addressed: • Because this measure was developed for use in the Merit-based Incentive Payment System (MIPS) and we believe that the information and testing provided should demonstrate that its use in MIPS will yield reliable and valid results and enable end users to make meaningful distinctions in the costs associated with the care provided to these patients. • The testing results provided, particularly for accountable-entity reliability, empirical validity, and the risk adjustment approach, do not provide the information needed to ensure that this measure produces the desired results: o It does not meet the minimum acceptable threshold of 0.7 for the accountable-entity reliability; o The empirical validity testing

does not include an assessment of this measure with a quality measure; o The current risk adjustment model does not adequately test and adjust for social risk factors; and o The testing provided in Section 2b4. Identification of Statistically Significant and Meaningful Differences in Performance does not directly address whether the costs attributed to physicians and practices enable us to distinguish low versus high performers.

Developer Response

Reliability Testing We would like to clarify that each of the measures has high reliability at both the TIN and TIN-NPI levels where the mean in fact exceeds 0.7, as shown in the testing materials. Specifically, the mean reliability for this measure for TINs is 0.78 and for TIN-NPIs is 0.72. This far exceeds the 0.4 mean reliability standard established through rulemaking for cost measures in MIPS. As noted by the commenter, testing results should demonstrate its reliability for use in MIPS; as such, the threshold set by CMS through regulatory processes is pertinent to any evaluation of reliability. Further, we note that the Scientific Methods Panel, whose role is to provide consistency and expertise in the scientific acceptability of measures, passed all measures on the reliability criterion. In fact, half of the SMP members rated the reliability as high, while the other half rated it as moderate (H: 4, M: 4). We reiterate that NQF does not set reliability thresholds, as stated in some of their materials, nor is there agreement in the literature on a threshold. Please see our submission materials and the CY 2022 PFS final rule (86 FR 65453 – 65454) for further details. Validity Results: Correlations with Related Quality Measures To clarify the commenter's concern about the lack of correlation analyses with quality measures, the empirical validity testing discussed during the Standing Committee evaluation meeting actually did include correlation analyses with quality measures. To recap that discussion for the commenter, we calculated the correlation between the cost measure and a MIPS quality measure for unplanned readmissions that we constructed using the public specifications. The results confirmed the expected relationship, namely that clinicians who have lower costs tend to have lower rates of unplanned readmissions, as demonstrated by the is high Pearson correlation between this cost measure and IP readmissions (0.57 at both the TIN and TIN-NPI levels) and unplanned readmissions (0.56 at the TIN level and 0.55 at the TIN-NPI level). SRF Testing Methodology To address the comment about the adequacy of SRF testing, we recap the discussion of testing during the Standing Committee evaluation meeting which included additional analyses which reflects guidance from organizations including NQF and ASPE about what considerations should be taken into account when assessing whether or not SRFs should be adjusted. We found that there is little impact on provider scores by risk adjusting for beneficiary dual status. We were however concerned that adjusting for dual status for this measure could risk masking providers' poor performance and exacerbate disparities in care because testing showed that providers who perform worse on dual beneficiaries perform worse on both dual and non-dual patients. That is, provider characteristics are more influencing the higher costs of episodes for patients with dual status, rather than patient factors. The testing approach that we discussed with the Standing Committee is one that has led to the decision to adjust for SRFs when results indicate that it is appropriate to do so. For example, the following two chronic condition measures that were finalized for MIPS 2022 do adjust for dual status: Diabetes and the Asthma/Chronic Obstructive Pulmonary Disease (COPD) episode-based cost measures. Finally, we agree with NQF's comments in the Draft Report that measures must be reviewed on a case-by-case basis to understand whether adjusting for SRFs is appropriate, to avoid unintended

consequences. Information in the cost measure meaningfully distinguishes between performance To confirm, the purpose of section 2b4 of the testing form is to demonstrate that there is clinically and practically significant variation in the measure scores. Given that testing results do show that this variation is present for the measure, they suggest that there are differences in performance, where some clinicians have low performance on the measure and some clinicians have high performance on the measure. We refer the commenter to other sections of the testing form to address the question of whether the costs included in the measure can meaningfully distinguish between high and low performance. Section 2b1 of the testing form describes how we convened a group of experts to provide detailed input on the measure specifications, including determining clinically related services that should be assigned to the measure. To gather a formal record of the workgroup's systematic input throughout development, workgroup members completed a face validity survey to assess the measure's ability to fulfill its intent to meaningfully compare and evaluate clinicians on cost efficiency. The results of the face validity vote showed that there was overall consensus agreement that the measure can distinguish good from poor performance. Finally, we share the commenter's interest in ensuring that end users can use the information from cost measures. Currently, MIPS participants receive patient-level episode-based cost measure reports which include the following information: episode identifiers (e.g., trigger date); list of all services rendered during the episode and the standardized costs, organized into service categories (e.g., post-trigger costs for outpatient facility costs); patient information (e.g., HCC risk score, sex). CMS will continue to consider feedback about what information is most useful for clinicians.

NQF Response

N/A

NQF Committee Response

Thank you for your comments. The Standing Committee recognizes that cost and resource use measures should be used in the context of and reported with quality measures. The Standing Committee discussed the relationship between cost and quality measures, emphasizing the importance of reporting performance to demonstrate improvements in cost while ensuring similar or higher levels of care quality. However, NQF criteria do not currently require that a cost measure be correlated with a quality measure. Rather, empirical validity testing should demonstrate that the measure's data elements are correct and/or the measure score correctly reflects the cost of care or resources provided. The Standing Committee further notes the need to ensure that providers serving people with SRFs are not penalized unfairly due to a lack of social risk adjustment. While the developer tested for social risk factors (SRFs) for the measure's risk adjustment model, some of the measures under review did not include these SRFs in the final model. Although the Standing Committee recognizes the importance of maximizing the predictive value of a risk adjustment model, elements of a risk model should be included or excluded based on a conceptual and empirical rationale. Thus, the Standing Committee considered the developer's empirical reliability and validity testing, including the Scientific Methods Panel's (SMP) input on both the reliability and validity testing, and the approach to the risk adjustment modeling and agreed to recommend these measures for endorsement.

National Quality Forum
1099 14th Street NW, Suite 500
Washington, DC 20005
http://www.qualityforum.org