NATIONAL QUALITY FORUM
Driving measurable health improvements together

Memo

October 27, 2022

To: Cost and Efficiency Standing Committee, Spring 2022

From: NQF staff

Re: Post-Comment web meeting to discuss NQF member and public comments received and NQF member expressions of support

Background

Healthcare cost measurement continues to be a critical component in assessing the efficiency of the U.S. healthcare system. Improving U.S. health system efficiency can simultaneously reduce cost and improve the quality of care provided. Measures in this portfolio are essential to evaluate the cost and efficiency of care and improve value through changes in care practices.

For the spring 2022, the Cost and Efficiency Standing Committee evaluated three new measures focused on condition-specific care episodes for 1) elective primary hip arthroplasty, 2) non-emergency coronary artery bypass graft, and 3) lumbar spine fusion for degenerative disease. The Standing Committee recommended all three measures for endorsement:

- 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)

Standing Committee Actions in Advance of the Meeting

1. Review this briefing memo and draft report.
2. Review and consider the full text of all comments received and the proposed responses to the post-evaluation comments (see Comment Brief).
3. Review the NQF members' expressions of support of the submitted measures.
4. Be prepared to provide feedback and input on proposed post-evaluation comment responses.

Comments Received

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 closed on September 26, 2022. Comments received by June 15, 2022 were shared with the Standing Committee prior to the measure evaluation meeting. 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 measures under review. This memo focuses on comments received after the Standing Committee's evaluation.

NQF members also had the opportunity to express their support ("support" or "do not support") for each measure submitted for endorsement consideration. One NQF member submitted an expression of non-support (Appendix A).

NQF staff have included all comments that were received (both pre- and post-evaluation) in this memo as the Comment Brief in Appendix B. The Comment Brief contains the commenter's name, comment, associated measure, and draft responses (including measure steward/developer responses, if appropriate) for the Standing Committee's consideration. Please review this memo and associated comments in advance of the spring 2022 post-comment meeting and consider the proposed responses for each comment.

In order to facilitate the discussion, the post-evaluation comments have been categorized into action items and major topic areas or themes. Although all comments are subject to discussion, the intent is not to discuss each individual comment during the post-comment call. Instead, NQF staff will spend the majority of the time considering the themes discussed below and the set of comments as a whole. Please note that the organization of the comments into major topic areas is not an attempt to limit the Standing Committee's discussion, and the Standing Committee can pull any comment for discussion. Measure stewards/developers were asked to respond to comments where appropriate. All developer responses along with the proposed draft Standing Committee responses have been provided in this memo and the Comment Brief.

Comments and Their Disposition

Themed Comments

Three major themes were identified in the post-evaluation comments, as follows:

1. Reliability Testing and Minimum Reliability Thresholds
2. Social Risk Adjustment
3. Cost and Quality Correlation

Theme 1 - Reliability Testing and Minimum Reliability Thresholds

The AMA voices concern with the testing results provided, specifically the accountable-entity reliability testing does not ensure that this measure will produce the desired results. The AMA voices concerns with the measure not meeting the minimum acceptable threshold of 0.7 for the accountable-entity reliability.

Measure Steward/Developer Response:

The developer has responded to the comment on reliability testing and minimum reliability thresholds and the full response can be found in Appendix B.

Proposed Standing Committee Response:

Thank you for your comment. The Standing Committee considered the Scientific Methods Panel's (SMP) input on both the reliability and validity testing, including the approach to the risk adjustment modeling and agreed to recommend these measures for endorsement.

Action Item:

Discuss and finalize Standing Committee response.

Theme 2 - Social Risk Adjustment

The AMA voices concern with the current risk adjustment model stating it 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.

Measure Steward/Developer Response:

The developer has responded to the comment on social risk adjustment and the full response can be found in Appendix B.

Proposed Standing Committee Response:

The Standing Committee acknowledges the commenter's concern. 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. The Standing Committee considered the developer's risk adjustment approach, including the Scientific Methods Panel's (SMP) input on validity testing, which was inclusive of the risk adjustment modeling approach, and agreed to recommend these measures for endorsement.

Action Item:

Discuss and finalize Standing Committee response.

Theme 3 - Cost and Quality Correlation

The AMA voices concern with the empirical validity testing not including an assessment of these measures with a quality measure.

Measure Steward/Developer Response:

The developer has responded to the comment on cost and quality correlation and the full response can be found in Appendix B.

Proposed Standing 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. Thus, the Standing Committee considered the developer's empirical validity testing, including the Scientific Methods Panel's (SMP) input on validity testing, and agreed to recommend these measures for endorsement.

Action Item:

Discuss and finalize Standing Committee response.

Appendix A: NQF Member Expression of Support Results

One NQF member provided their expressions of support/do not support. None of the measures under consideration received support. Results for each measure are provided below.

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

Member Council	Commenter Names, Organizations	Support	Do Not Support	Total
Health Professional (HPR)	Koryn Rubin, American Medical Association	0	1	1
Total	*	0	1	1

*Cell intentionally left blank.

NQF #3625 Non-Emergent Coronary Artery Bypass Graft (CABG) Measure (Centers for Medicare & Medicaid Services/Acumen, LLC)

Member Council	Commenter Names, Organizations	Support	Do Not Support	Total
Health Professional (HPR)	Koryn Rubin, American Medical Association	0	1	1
Total	*	0	1	1

*Cell intentionally left blank.

NQF #3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure (Centers for Medicare & Medicaid Services/Acumen, LLC)

Member Council	Commenter Names, Organizations	Support	Do Not Support	Total
Health Professional (HPR)	Koryn Rubin, American Medical Association	0	1	1
Total	*	0	1	1

*Cell intentionally left blank.

Appendix B: Comment Brief

Post-Evaluation Measure-Specific Comments on Cost and Efficiency Spring 2022 Submissions

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/neurosurgeons 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
Thank you for your comment. It has been shared with the Standing Committee and the measure developer.

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

Thank you for your comment. It has been shared with the Standing Committee and the measure developer.

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 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

Thank you for your comment. It has been shared with the Standing Committee and the measure developer.

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.

Public Comments on Cost and Efficiency Spring 2022 Draft Report

No comments were received regarding the draft technical report.

Pre-Evaluation Measure-Specific Comments on Cost and Efficiency Spring 2022 Submissions

NQF #3623 Elective Primary Hip Arthroplasty Measure (Recommended)

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

Comment ID#: 8108 (Submitted: 06/14/2022)

Council / Public: HPR

Level of Support: Member Does NOT Support

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) Measure (Recommended)

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

Comment ID#: 8109 (Submitted: 06/14/2022)

Council / Public: HPR

Level of Support: Member Does NOT Support

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 Measure (Recommended)

Ms. Koryn Y. Rubin, MHA, American Medical Association
Comment ID#: 8110 (Submitted: 06/14/2022)

Council / Public: HPR

Level of Support: Member Does NOT Support

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.