NATIONAL QUALITY FORUM
Driving measurable health improvements together

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3626

De.2. Measure Title: Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services

De.3. Brief Description of Measure: 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.

IM.1.1. Developer Rationale: Lumbar spine fusion surgeries comprise some of the largest admission expenditures in the Medicare program, and are increasingly prevalent among Medicare patients. Total admission expenditures for these procedures exceeded $3.6 billion in 2013, and more than 6 million Medicare patients were diagnosed with lumbar degenerative conditions between 2006 and 2012. [1][2] Currently, there are substantial opportunities to improve the cost-efficiency and quality of care related to these procedures, given the high variation in treatment options. Primarily these include the use of less-invasive surgical techniques to reduce post-operative complications. [3]

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode-based cost measure (also referred to in this form as "the Lumbar Spine Fusion" measure) was recommended for development by an expert clinician committee (the Musculoskeletal Disease Management - Spine Clinical Subcommittee, composed of 22 clinician experts affiliated with 19 specialty societies) because of its impact in terms of Medicare cost and patient populations, as well as the opportunity for incentivizing cost-effective, high quality care in this area. Based on the initial recommendations from the Clinical Subcommittee, a subsequent Lumbar Spine Fusion clinician expert workgroup (composed of 13 members affiliated with 13 specialty societies) provided extensive, detailed input on this measure. Workgroup input has helped ensure the measure's ability to fairly evaluate clinician cost performance for Lumbar Spine Fusion surgeries and to promote efficient and high quality care for Medicare patients undergoing these procedures.

[1] Zorica Buser et al., "Spine Degenerative Conditions and Their Treatments: National Trends in the United States of America." [In eng]. Global Spine J 8, no. 1 (Feb 2018): 57-67.

[2] Steven D. Culler et al., "Incremental Hospital Cost and Length-of-Stay Associated with Treating Adverse Events among Medicare Beneficiaries Undergoing Lumbar Spinal Fusion During Fiscal Year 2013." [In eng]. Spine (Phila Pa 1976) 41, no. 20 (Oct 15, 2016): 1613-20.

[3] Christina L. Goldstein et al., "Comparative Outcomes of Minimally Invasive Surgery for Posterior Lumbar Fusion," Clinical Orthopaedics and Related Research," 472, no.6 (2014): 1727-1737, https://doi.org/10.1007/s11999-014-3465-5.

De.1. Measure Type: Cost/Resource Use

S.5. Data Source: Claims

S.3. Level of Analysis: Clinician: Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. High impact or high resource use:
The measure focus addresses:
– a demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality).
AND
1b. Opportunity for Improvement:
Demonstration of resource use or cost problems and opportunity for improvement, i.e., data demonstrating
considerable variation cost or resource across providers

[Response Begins]

1a. High Impact or high resource use.

- The developer cites that more than six million Medicare patients were diagnosed with lumbar degenerative conditions between 2006 and 2012, and that the total admission expenditures for lumbar spine fusion surgeries exceeded $3.6 billion in 2013.

- The developer posits that there are opportunities to improve the cost and quality of care related to these procedures, namely using less-invasive surgical techniques to reduce post-operative complications.
- This measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo surgery for lumbar spine fusion during the defined performance period.

1b. Opportunity for Improvement:

- The developer provides a distribution of performance scores for clinician groups (identified by Tax Identification Number [TIN]) and individual clinicians (identified by a combination of TIN and National Provider Identifier [NPI]) attributed 10 or more Lumbar Spine Fusion episodes from January 1, 2019, to December 31, 2019.
- These scores reflect 1,415 clinician group practices and 3,330 individual practitioners, corresponding to 54,768 episodes of care for 54,768 beneficiaries. Episodes are included from all 50 States and D.C. in the following settings: acute IP hospitals, OP facilities, ambulatory/office-based care centers, and ambulatory surgical centers (ASC).
- For the TIN-level scores, the mean was 1.01 with an interquartile range (IQR) of 0.10. For the TIN-NPI-level, the mean score was 1.00 with an IQR or 0.11.

Questions for the Committee:

- *Has the developer demonstrated this is high impact, high-resource use area to measure?*
- *Is there a sufficient variation in performance across hospitals that warrants a national performance measure?*

[Response Ends]

Staff preliminary rating for opportunity for improvement: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Committee Pre-evaluation Comments:

1a. Evidence

- Yes

1b. Gap in Care/Opportunity for Improvement and Disparities

- Yes

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

2b. Validity: Alignment of Specifications with Intent (includes threats to validity [e.g., attribution, costing method, missing data]) Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Multiple Data Sources; and Disparities.

Measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators: NQF Scientific Methods Panel (SMP)

R: H-4, M-4, L-0, I-0

V: H-0, M-6, L-2, I-0

Measure evaluated by Technical Expert Panel? ☐ **Yes** ☒ **No**

Evaluators: N/A

Reliability

2a1. Specifications:

The measure is well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability. All measures that use the ICD classification system must use ICD-10-CM.

2a2. Reliability testing:

Demonstration that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

2a1. Specifications

- The cost measure is calculated as the sum of the ratio of observed to expected payment-standardized cost to Medicare for all Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes attributed to a clinician or clinician group. The resulting average episode cost ratio is then multiplied by the national average observed episode cost to generate a dollar figure.
- The episode window spans from 30 days prior to the trigger day through 90 days after, and includes costs from certain clinically-related services from Medicare Parts A and B claims during the episode window.
- Costs are standardized to account for differences in Medicare payments for the same service(s) across Medicare providers.
- One SMP member questioned why the measure is attributed to co-surgeons, but not other members of the surgical team (e.g., anesthesiologist). Additionally, this SMP member also questioned why skilled nursing facility claims were not included.
  - The developer provided responses to this member's comments. Specifically, the developer noted that the attribution methodology focuses on the clinician(s) performing the lumbar spine fusion procedure by attributing an episode to the clinician(s) who bill the trigger code (CPT/HCPCS procedure code). This can be both the main and assistant clinician. We use this methodology as the measure intent is to assess costs related to the role of the

clinician performing the surgical procedure. Since the role of an anesthesiologist or CRNA is distinct from performing the surgery itself, this measure does not attribute episodes to members of the care team who do not bill the trigger procedure.

- Additionally, the developer noted that the measure includes SNF costs where the SNF claim's qualifying inpatient stay is the same as the trigger inpatient procedure. This ensures that SNF is only assigned to an episode where it is closely related to the inpatient surgical procedure. This is detailed in the MIF, Section A.3.

2a2. Reliability Testing:

- The developer used a signal-to-noise analysis to evaluate reliability at the group practice (TIN) and individual clinical (TIN-NPI) levels using a split-sample method, calculated from a larger sample of episodes in 2018 and 2019 to get enough volume per TIN and TIN-NPI (with minimum of 10 episodes per TIN and TIN-NPI). The developer calculated Shrout-Fleiss intraclass correlation coefficients (ICCs).
- The mean signal-to-noise reliability was 0.78 for TINs and 0.72 for TIN-NPIs. Reliability was slight lower at the 10th and 25th deciles (0.64 and 0.69, respectively at TIN and 0.60 and 0.65 at TIN-NPI) and higher at the 90th percentile (0.92 TIN and 0.84 TIN-NPI). Reliability at the practice size was also evaluated, with the average reliability scores increasing from 0.71 (1 clinician) to 0.95 (21+ clinicians) for TINs. Pearson correlation and ICC coefficients between the split-sample measures scores were 0.73 at the TIN-level and 0.67 at the TIN-NPI level.
- The SMP did not raise any major concerns and passed the measure on reliability (H-4, M-4, L-0, I-0).

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *Do you have any concerns with the reliability testing that was not identified by the Scientific Methods Panel?*

[Response Ends]

Staff Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

2a1. Reliability-Specifications

- No concerns

2a2. Reliability-Testing

- No concerns

Validity

2b1. Specifications align with measure intent:

The measure specifications are consistent with the measure intent and captures the most inclusive target population.

2b2. Validity Testing:

Demonstration that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided.

2b3. Exclusions:

Exclusions are supported by the clinical evidence**,** AND/OR There is a rationale or analysis demonstrating that the measure results are sufficiently distorted due to the magnitude and/or frequency of then on-clinical exclusions; AND Measure specifications for scoring include computing exclusions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); AND If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

2b4. Risk Adjustment:

For resource use measures and other measures when indicated: an evidence-based risk-adjustment strategy is specified and is based on patient factors (including clinical and sociodemographic risk factors) that influence the measured outcome and are present at start of care, and has demonstrated adequate discrimination and calibration, OR rationale/data support no risk-adjustment/-stratification.

2b5. Meaningful Differences:

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

2b6. Multiple Data Sources:

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

2c. Disparities: If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender), OR rationale rationale/data justifies why stratification is not necessary or not feasible.

2b1. Specifications Align with Measure Intent:

- Attribution:
  - This measure attributes lumbar spine fusion episodes to clinicians (TIN-NPIs) billing the triggering procedure code. At the clinician group level, an episode is attributed to the TIN if its TIN-NPI(s) attributed an episode by billing the triggering procedure and all episodes across the TIN's NPI(s) are aggregated. If the same episode is attributed to more than one NPI within a TIN, this episode is only attributed to the TIN once.

- Costing approach:
  - The developer notes that this measure uses Medicare Standardized Pricing. The methodology used to payment standardize the Medicare claims used to specify this measure is available for download ("CMS Price (Payment) Standardization") from the following URL: https://www.resdac.org/articles/cms-price-payment-standardization-overview.

2b2. Validity Testing:

- The SMP passed this measure on validity (H-0, M-6, L-2, I-0).
- Face Validity
  - The developer gathering input for the measure from clinician experts and other stakeholders during measure development.
  - The clinical subcommittee included 22 members with relevant clinical experience and a Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Clinician Expert Workgroup ("workgroup") of 13 members.
  - Workgroup members (9/13 members voted - 69% response rate) agreed that the measure could accurately capture a clinician's risk adjusted cost to Medicare for patients who receive Lumbar Spine Fusion, with mean ratings of 3.9 or higher out of scale of 6 for 5 face validity questions related to triggers, exclusions, service assignment, episode window identification, and risk adjustment variables (mean response ratings 4.5 on all 5 questions or somewhat to moderately agree). The developer was unable to obtain a mean rating on the question "The scores obtained from the Non-Emergent Lumbar Spine Fusion measure as specified will provide an accurate reflection of the costs for episodes of care, and can be used to distinguish good and poor performance on cost effectiveness."
- Empirical Validity
  - The empirical validity was evaluated by examining correlation with the Medicare Spending Per Beneficiary (MSPB) Hospital Measure (NQF# 2158), which assesses the risk-adjusted cost to Medicare for services performed by hospitals and other healthcare providers during an MSPB-Hospital episode.
  - Specifically, the developer analyzed the distribution of Lumbar Spine Fusion measure scores (i.e., observed to expected cost [O/E] ratios) across MSPB performance ratings.
  - Empirical testing shows the mean cost scores (O/E ratios) were highest for TINs with lowest performance on the MSPB Hospital Measure (low cost efficiency) at 1.04, decreasing as performance ratings increased to 0.96 at performance rating from 5-10 (best cost efficiency), as expected. A similar result for TIN/NPI with mean cost score 1.04 for lowest performance rating to 0.96 at highest performance rating.

2b3. Clinical Inclusions and Exclusions/Evidence to Support Clinical Logic

- The developer excludes certain episodes (e.g., patients with cancer, patients with an infection, patients the underwent a redo lumbar fusion) to achieve fair comparisons across providers.
- The developer reports that the statistical results of the exclusions provide evidence that excluded episodes are not comparable to the overall measure population.

2b4. Risk adjustment

- The developer includes 122 risk factors in the overall risk model.
- The risk model was informed by covariates recommended by developer-convened expert panel and the CMS Hierarchical condition categories (HCC), as well as demographic information from the Medicare enrollment file (e.g., age, race, disability, dual status). Information on income, education, and unemployment were obtained from Census American Community Survey data.
- The risk adjustment model was performed separately for three measure sub-groups based on level of fusion: (i) One Level Lumbar Spine Fusion; (ii) Two Level Lumbar Spine Fusion; and (iii)

Three Level Lumbar Spine Fusion. The social risk factors were included AFTER the base risk adjustment (for clinical factors).

- Stepwise regression was used to include sex, dual status, sex+dual status, sex + dual status + race, sex + dual status + income + education + unemployment, sec + dual status + Agency for Healthcare Research & Quality Socioeconomic Status (SES) Index, sex + dual status + income + education + unemployment + race, and sex + dual status + race + AHRQ SES Index.
    - The developers report that the analyses found the relationship between the various social risk factors tested and the measure cost scores were inconsistent across factors and sometimes negative. The developer reports that including these factors could introduce bias into the measure.
    - Many significant p values indicate social risk factors are predictive of resource use. However, analysis results suggested that adding social risk factors to the measure risk-adjustment model had minimal impact on measure performance and was largely redundant with current model prediction.
    - This was determined using two methods: by (i) analyzing differences in percentiles of observed to expected episode cost (O/E) ratios both with and without social risk factors in the model, and (ii) examining correlations between measure scores calculated with and without social factors.
    - Both of these tests demonstrated a minimal impact on performance – even for providers at high and low extremes of risk - from including social risk factors in the model.
- The overall R-squared for the measure was 0.516 (and adjusted R-squared 0.513). The average observed to expected cost was generally close to one, 0.99 to 1.01, across risk deciles, and the average O/E cost ratios for all risk deciles are close to 1.0.

2b5: Meaningful Differences

- For TINs, the standard deviation is 0.09, and 99/1, 90/10, and 75/25 percentile ratios are 1.56, 1.23, and 1.10, respectively.
- For TIN-NPIs, the standard deviation is 0.10, and 99/1, 90/10, and 75/25 percentile ratios are 1.56, 1.24, and 1.11, respectively.
- Scores were not influenced by region or number of episodes performed.

2b6. Multiple Data Sources

- N/A – this measure used Medicare administrative claims

2c. Disparities

- The developer did not provide performance data distributions by sub-populations.

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., correlations, exclusions, risk-adjustment approach, etc.)?*
- *Does the SC have any concerns related to the risk adjustment model (e.g., the r-squared values, lack of social risk factor adjustment)*

[Response Ends]

Staff preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

2b1. Validity -Testing

- No concerns

2b2. Additional threats to validity

- No concerns

2b3. Additional Threats to Validity

- No concerns

2b4/2c. Additional Threats to Validity: Risk Adjustment

- No concerns

2b5. Threats to Validity: Meaningful Differences

- No concerns

2b6. Threats to Validity: Missing Data/Carve Outs

- No concerns

Criterion 3. Feasibility

3. Feasibility

The extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure uses administrative claims data. Data are generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition.
- Data is coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims).
- The developer indicates that all data elements for this measure are in defined fields in a combination of electronic claims.
- The developer does not indicate that there are any fees associated with the use of this measure.

Questions for the Committee:

- *Are there any concerns regarding feasibility?*

Staff preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

3. Feasibility

- No concerns

Criterion 4: Usability and Use

Use

4a. Use. evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency.

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4a.2. Feedback on the measure by those being measured or others.

Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

4a1. Current uses of the measure

- **Publicly reported? ☒ Yes ☐ No**
- **Current use in an accountability program? ☒ Yes ☐ No ☐ UNCLEAR**

Accountability program details

- The developer indicated that the measure is currently used in the Quality Payment Program (QPP) Merit-based Incentive Program System (MIPS).
- As specified in the CY 2020 Physician Fee Schedule final rule (84 FR 62959 through 62979), this measure has been implemented as part of Merit-based Incentive Payment System (MIPS) beginning in the 2020 MIPS performance year and 2022 MIPS payment year.

4a2.Feedback on the measure by those being measured or others

- During the development of this measure, the Lumbar Spine Fusion Field Test Reports were provided to a sample of eligible clinician groups and clinicians. Each report included information for the Lumbar Spine Fusion measure if the clinician or clinician group was attributed 10 or more episodes. All stakeholders, including those who did not receive a field test report, could review a mock field test report that was posted on the CMS website.
- During field testing, the developer conducted education and outreach activities, including a national webinar, office hours with specialty societies, and Help Desk support. The developer sought feedback on the reports and measure specifications through an online survey, with the option to attach a comment letter.

- After completing field testing, the developer compiled the feedback provided through the survey and comment letters into a measure-specific report, which was then provided to the Lumbar Spine Fusion Clinician Expert workgroup, along with empirical analyses to inform their discussion and evaluation of any refinements needed to ensure that the measure is capturing what it was intended to capture.
- Stakeholders provided cross-cutting feedback on risk adjustment variables (e.g., cognitive and functional status, academic medical centers, and socioeconomic status), attribution methodology, episode windows and assigned services, and alignment with cost and quality.

Additional Feedback:

- The Lumbar Spine Fusion measure was implemented in MIPS after going through the pre-rulemaking process and notice-and-comment rulemaking.
- The measure was submitted to and included in the 2018 Measures Under Consideration (MUC) List. It was then considered by National Quality Forum (NQF)'s Measure Applications Partnership (MAP) Clinician Workgroup and Coordinating Committee in December 2018 and January 2019, respectively.
- The MAP voted to conditionally support this measure for rulemaking, conditional on submission to the NQF review and endorsement process.
- The MAP noted that CMS and the Cost and Efficiency Standing Committee should continue to evaluate the risk adjustment model of this measure and consider whether there is need to account for social risk factors in the model.
- The MAP also noted that review of the measure should ensure an appropriate attribution methodology and that the measure adequately considers the issue of small numbers.
- The MAP noted that cost measures should continue surveillance for unintended consequences such as stinting of care and reduced quality of care, and that cost measures should be paired with balancing measures (e.g., quality, efficiency, access, and appropriate use measures) as one way to safeguard against these issues.
- The MAP recognized a need for continuous feedback and testing of measures as they are implemented. Lastly, the MAP noted a need to provide greater education on these measures as well as for greater transparency of the measure specifications and testing results.

Questions for the Committee:

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *How has the measure been vetted in real-world settings by those being measured or others?*

[Response Ends]

Staff preliminary rating for Use: ☒ Pass ☐ No Pass

Committee Pre-evaluation Comments:

4a1. Use

- No concerns

4a2. Usability

- No concerns

Usability

4b. Usability.

The extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement.

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

4b2. Benefits vs. harms.

Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b1. Improvement results

- The developer did not report results over time, as this measure has not been used prior to 2020.
- From the "Opportunity for Improvement" section:
    - The developer provides a distribution of performance scores for clinician groups (TIN) and individual clinicians (TIN-NPI) attributed 10 or more Lumbar Spine Fusion episodes from January 1, 2019, to December 31, 2019.
    - For the TIN-level scores, the mean was 1.01 with an interquartile range (IQR) of 0.10. For the TIN-NPI-level, the mean score was 1.00 with an IQR or 0.11.

4b2. Unintended consequences

- The developer notes that there were no unintended consequences to individuals or populations identified during the development and testing of this measure.

4b2.Potential harms

- No potential harms were identified.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *What benefits, potential harms or unintended consequences should be considered?*

[Response Ends]

Staff preliminary rating for Usability and Use: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

4a1. Use - Accountability and Transparency

- High number eligible clinicians. Currently used in QPP MIPs program
- It's not clear by the information provided whether this measure is currently publicly reported or not.

4a2. Use - Feedback on the measure

- Field testing done by Acumen and CMS in national field test of 11 episode based cost measures
- No concerns

4b1. Usability – Improvement

- Newly implemented measure and no data to demonstrate improvement
- No concerns

4b2. Usability – Benefits vs. harms

- no unintended consequences to individuals or populations have been identified during testing and development of this measure.
- There is the potential harm in that accountability for costs in the short term may influence decision making that may have an impact beyond the episode of measurement.

Criterion 5: Related and Competing Measures

- The developers identified that there are no competing NQF-endorsed or non-NQF-endorsed cost measures with the same measure focus and/or target population submitted to NQF or implemented in MIPS.

Harmonization

- N/A

Committee Pre-evaluation Comments:

5. Related and Competing

- No concerns

Public and Member Comments (Submitted as of June 15, 2022)

Member Expression of Support

- One NQF member submitted an expression of "do not support" for the measure.

Comments

Comment 1 by: Submitted by Koryn Rubin, American Medical Association

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.

Combined Scientific Methods Panel Preliminary Analysis of Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3626

Measure Title: Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☒ **No**

   Submission document: "MIF_3626" document, items S.1-S.22

   NOTE*: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

2. **Briefly summarize any concerns about the measure specifications.**

   Panel Member 1: How easy/difficult it would be to determine clinically relevant services associated with the lumber surgery?

   Panel Member 4: No concerns

   Panel Member 5: None

Panel Member 6: NA

Panel Member 8: The role of the sub-groups within the measure is not clear. Since there is no testing results for sub-groups, I'm assuming these are merely descriptive. Also, I am not clear on why the measure is attributed to co-surgeons, but not other members of the surgical team (e.g., anesthesiologist). I also noticed the SNF claims are not included - this seems like an important omission.

RELIABILITY: TESTING

Type of measure:

☐ **Outcome (including PRO-PM)** ☐ **Intermediate Clinical Outcome** ☐ **Process**

☐ **Structure** ☐ **Composite** ☒ **Cost/Resource Use** ☐ **Efficiency**

Data Source:

☐ **Abstracted from Paper Records** ☒ **Claims** ☐ **Registry**
☐ **Abstracted from Electronic Health Record (EHR)** ☐ **eMeasure (HQMF) implemented in EHRs**
☐ **Instrument-Based Data** ☒ **Enrollment Data** ☒ **Other (please specify)**
Panel Member 1: Long-term Minimum Data Set, and Common Medicare Environment
Panel Member 2: CMS administrative data sets
Panel Member 3: Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment
Panel Member 5: Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment
Panel Member 6: Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment
Panel Member 8: Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment

Level of Analysis:

☒ **Individual Clinician** ☒ **Group/Practice** ☐ **Hospital/Facility/Agency** ☐ **Health Plan**
☐ **Population: Regional, State, Community, County or City** ☐ **Accountable Care Organization**
☐ **Integrated Delivery System** ☐ **Other (please specify)**

Measure is:

☒ **New** ☐ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

Submission document: "MIF_3626" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☒ **Measure score** ☐ **Data element** ☐ **Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of **patient-level data** conducted?

   ☐ **Yes** ☒ **No**
6. **Assess the method(s) used for reliability testing**

Submission document: Testing attachment, section 2a2.2

Panel Member 1: Signal-to-noise and split- sample reliability tests, which are considered appropriate for reliability testing.

Panel Member 2: S/N, split sample

Panel Member 4: The methods were appropriate -- the developer used a STN and test-retest approach using split sampling.

Panel Member 5: Reasonable, split-sample approach

Panel Member 6: Split sample reliability testing

Panel Member 8: Developers use a reliability scores (e.g., Adams) and a 'split' sample with correlation. Seems appropriate.

Panel Member 9: Developer used signal-to-noise analysis to evaluate reliability at the group practice (TIN) and individual clinical (TIN/NPI) levels using split sample, calculated from a larger sample of episodes in 2018 and 2019 to get enough volume per TIN and TIN/NPI (with minimum of 10 episodes per TIN and TIN/NPI). They calculated Shrout-Fleiss intraclass correlation coefficients (ICCs).

7. **Assess the results of reliability testing**

Submission document: Testing attachment, section 2a2.3

Panel Member 1: Signal-to-noise: The average measure reliability was estimated to be 0.78 for TINs and 0.72 for TIN-NPIs. Reliability for groups of different practice sizes was high, with mean reliability for the smallest TINs at 0.71. Split-sample reliability test: The ICC coefficient was 0.73 for TINs and 0.67 for TIN-NPIs, indicating high or moderate overall reliability for TINs and TIN-NPIs.

Panel Member 2: S/N: median and mean scores are adequate. Those for smallest practices a little low split sample: adequate, again TIN-NPI at 0.67 lower

Panel Member 4: The measure reported at TIN and TIN-NPI levels was reliable, regardless of the reliability testing method (median of .79 and 0.71 for TIN and TIN-NPI respectively). The IQR of 0.69 - 0.87 for TIN and .65 - 0.79 for TIN-NPI is also sufficient. The only caveat is that the developer did the testing using a 10 episode threshold which is not part of the specification.

Panel Member 5: The mean reliability for TINs is 0.78 and for TIN-NPIs is 0.72. The average reliability scores increase from 0.71 (1 clinician) to 0.95 (21+ clinicians) for TINs. The ICC coefficient was 0.73 at the TIN-level, and 0.67 at the TIN-NPI level

Panel Member 6: ICC (2,1) is median of 0.79 at TIN level and 0.71 at TIN-NPI level.

Panel Member 8: My primary concern in TIN-NPIs with low volume - this version of the measure is reported to have a median reliability score of 0.71. However, the 10th percentile is 0.60, suggesting there is a cluster of surgeons with low volumes and low reliability. The importance of volume is demonstrated in Table 2. The split sample correlation for TIN-NPIs is 0.67 (Table 3), again relatively low given the large sample size.

Panel Member 9: Mean reliability was 0.78 for TINs and 0.72 for TIN/NPIs. Reliability was slight lower at the 10th and 25th deciles (i.e., 0.64 and 0.69 respectively at TIN and 0.60 and 0.65 at TIN/NPI) and much higher at the 90th percentile (i.e., 0.92 TIN and 0.84 TIN/NPI). Reliability at the practice size was also evaluated, with the average reliability scores increasing from 0.71 (1 clinician) to 0.95 (21+ clinicians) for TINs. Pearson correlation and ICC coefficients between the split-sample measures scores were 0.73 at the TIN-level and 0.67 at the TIN-NPI level.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable** (score-level testing was not performed)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☒ **No**

   ☒ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and **all** testing results):

    ☒ **High** (NOTE: Can be HIGH **only if** score-level testing has been conducted)

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has **not** been conducted)

    ☐ **Low** (NOTE: Should rate **LOW** if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate **INSUFFICIENT** if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

    Panel Member 1: Both the signal-to-noise and split-sample analyses indicate high level of reliability.

    Panel Member 2: S/N, split sample averages adequate.

    Panel Member 4: Developer showed adequate reliability but limited sample to 10 or more episodes, which is not consistent with the specification.

    Panel Member 5: Split-half, with moderate findings was a reasonable approach

    Panel Member 8: The measure probably has high reliability at the group level where volumes are higher, but moderate reliability at the TIN-NPI level. Given my concerns about low volume providers, I am rating this moderate.

    Panel Member 9: Overall, testing results indicated high measure score reliability with an average of 0.84 for TINs and 0.75 for TIN-NPIs at a volume threshold of 10 episodes. Reliability for groups of different practice sizes was also high, with mean reliability for the smallest TINs at 0.76. The split-sample reliability analysis also shows evidence of reliability and repeatability of the performance measure with ICC coefficient 0.80 for TINs and 0.64 for TIN-NPIs, indicating moderate to high overall reliability for TINs and TIN-NPIs.

VALIDITY: TESTING

12. **Validity testing level:** ☒ **Measure score** ☒ **Data element** ☐ **Both**

13. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?** *NOTE that data element validation from the literature is acceptable.*

Submission document: *Testing attachment, section 2b1.*

☒ **Yes**

☐ **No**

☒ **Not applicable** (data element testing was not performed)

14. **Method of establishing validity of the measure score:**

☒ **Face validity**

☒ **Empirical validity testing of the measure score**

☐ **N/A (score-level testing not conducted)**

15. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☒ **No**

☐ **Not applicable** (score-level testing was not performed)

16. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

Panel Member 1: Face validity of this measure was assessed through a structured process for gathering detailed input from clinician experts and other stakeholders including (i) the Musculoskeletal Disease Management - Spine Clinical Subcommittee, (ii) the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels workgroup, (iii) a Technical Expert Panel (TEP), (iv) a Person and Family Committee (PFC), and (v) stakeholder feedback from national field testing. Empiric validity was evaluated by examining correlation with an NQF endorsed measure of resource use: the Medicare Spending Per Beneficiary (MSPB) Hospital Measure (NQF# 2158)

Panel Member 2: Face validity: TEP Empirical validity: correlation with MSPB hospital

Panel Member 4: Developer used face validity and correlation between Lumbar measure and MSPB measure.

Panel Member 5: TEP and correlation with other measures.

Panel Member 6: Testing includes comparison of hospital MSPB for all cases to episode cost for selected cases - see #3623 for concerns.

Panel Member 8: I appreciate the detail on how clinicians provided guidance and feedback throughout the measure development process. I also like the selection of the use of the MSPB as a comparator.

Panel Member 9: The Lumbar Spine Fusion measure underwent a structured process for gathering detailed input from clinician experts and other stakeholders during measure development. Ther Clinical subcommittee included 22 members with relevant clinical experience and a Lumbar Spine Fusion workgroup (TEP) of 13 members. The empirical validity was evaluated by examining correlation with the Medicare Spending Per Beneficiary (MSPB) Hospital Measure (NQF# 2158), which assesses the risk-adjusted cost to Medicare for services performed by hospitals and other healthcare providers during an MSPB-Hospital episode. Specifically, they analyzed the distribution of Lumbar Spine Fusion measure scores (i.e., observed to expected cost [O/E] ratios) across MSPB performance ratings.

17. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    Panel Member 1: The face validity of the measure was established with the fact that out of 9 respondents to the survey, substantial majorities (6 to 8 respondents) agreed that each of the measure specifications helps the measure capture clinician cost performance as intended, and that the scores from the measure, as currently specified, provide an accurate reflection of clinician cost effectiveness. This was in addition to the overall mean response rating of 4.5 out of 6, indicating a fair level of agreement with each of the key measure components. Empirical Validity was established by the negative association with the MSPB measure.

    Panel Member 2: Face validity: TEP generally in agreement. Biggest area of disagreement was service assignment and committee should seek more info on this. Empirical validity: Correlation with MSPB in right direction but modest.

    Panel Member 4: Face validity was sufficient, lumbar measure showed weak but expected direction correlation with MSPB.

    Panel Member 5: Reasonable results. The mean rating from these five questions indicates overall consensus agreement on the measure specifications

    Panel Member 6: Even if methodology was sound, there is not enough variability between the SMPB performance rating categories to show meaningful information. (same concern as #3623).

    Panel Member 8: The face validity results provide good insights into how difficult it is to build this type of resource use measure. I appreciate the survey results, but would have really like the overall rating on 'will provide an accurate reflection of the costs'. In terms of the empirical testing, the episode cost follow the expected pattern. This is a good start, but additional aspects of validity (e.g., predictive) should be tested over time.

    Panel Member 9: TEP members (appears 9/13 members voted 69% response rate) agreed that the measure could accurately capture a clinician's risk adjusted cost to Medicare for patients who receive Lumbar Spine Fusion, with mean ratings of 3.9 or higher out of scale of 6 for 5 face validity questions related to triggers, exclusions, service assignment, episode window identification, and risk adjustment variables (mean response ratings 4.5 on all 5 questions or somewhat to moderately agree). They were unable to obtain a mean rating on the question "The scores obtained from the Non-Emergent Lumbar Spine Fusion measure as specified will provide an accurate reflection of the costs for episodes of care, and can be used to distinguish good and poor performance on cost effectiveness." Empirical testing show the mean cost scores (observed/expected ratio) were highest for TINs with lowest performance on the MSPB Hospital Measure (low cost efficiency) at 1.04, decreasing as performance ratings increased to 0.96 at performance rating from 5-10 (best cost efficiency). A similar result for TIN/NPI with mean cost score 1.04 for lowest performance rating to 0.96 at highest performance rating. This does not show a significantly wide range of difference in efficiency (diff of 0.08) in cost ratios between low and high performing clinicians/groups.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

18. **Please describe any concerns you have with measure exclusions.**

    Submission document: Testing attachment, section 2b2.

    Panel Member 1: Similar question as for non-emergent CABG measure – if a patient dies within the episode duration due to mismanagement (aka, low-quality care), should or shouldn't the patient be excluded from the measure?

Panel Member 2: No concerns about beneficiary exclusions. As noted above, we do not have adequate documentation of how costs are narrowly defined to only be those associated with Lumbar Fusion and its post-acute care. TEP vote on this issue, service assignment, saw greatest split 6-3

Panel Member 4: Defer to standing committee on exclusions

Panel Member 5: None. 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.

Panel Member 6: NA

Panel Member 8: It would be helpful in Table 6 to show the number of cases dropped with each exclusion. The exclusion represent real surgical cases, many of which include serious opportunity for care improvement. No doubt, dropping these cases improves the measure performance on reliability. However, the 'cost' of these exclusions is a less accurate picture of reality. The exclusions seem to go too far.

Panel Member 9: Exclusions seem appropriate.

19. **Risk Adjustment**

Submission Document: Testing attachment, section 2b3

19a. **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☒ **Stratification**

19b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒ Not applicable

19c. **Social risk adjustment:**

19c.1 Are social risk factors included in risk model? ☒ Yes ☒ No ☐ Not applicable

19c.2 Conceptual rationale for social risk factors included? ☒ Yes ☐ No

19c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☐ No

19d. **Risk adjustment summary:**

19d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No
19d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☐ No
19d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No
19d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☐ No
19d.5. Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☐ No

19e**. Assess the risk-adjustment approach**

Panel Member 1: I don't agree with the developer not including SRFs even after clearly observing that at least dual status is consistently significant across all the models considered.

Panel Member 2: R-square of 0.51 high. The social risk adjustment decisions need more information and need to be reviewed. The issues I raised in CABG apply here: there is a dismissal of the variables because when entered as a group some signs are in the wrong direction, but this often happens when correlated measures are added to a regression model. We need analysis of the impact on prediction and aggregate direction. Also, while 95% of TINs and TIN-NPIs move less than 5 percentiles, we need data on movement of providers that have large panels of social disadvantaged patients.

Panel Member 4: CMS HCC is the basis of risk adjustment for this measure and includes a wide range of conditions and interaction terms (120+). Potential for overfit.

Panel Member 5: Reasonable approach. To analyze the validity of the current risk adjustment model, we examined three analyses: (a) R-squared and adjusted R-squared for the regression models, (b) predictive ratios to examine the fit of the models at different levels of patient complexity, and (c) coefficient estimates, standard errors, and p-values for the risk-adjustment model.

Panel Member 6: Typical CMS statistical risk adjustment using HCCs and demographic factors.

Panel Member 8: Impressive work building up the risk model.

Panel Member 9: 122 risk factors / 3 risk categories; The risk model was informed by covariates recommended by expert panel and the CMS HCC categories, as well as demographic information from the Medicare enrollment file (e.g., age, race, disability, dual status). Information on income, education, and unemployment were obtained from Census ACS data. The risk adjustment model was performed separately for 3 measure sub-groups based on level of fusion: (i) One Level Lumbar Spine Fusion; (ii) Two Level Lumbar Spine Fusion; and (iii) Three Level Lumbar Spine Fusion. The social risk factors were included AFTER the base risk adjustment (for clinical factors). Stepwise regression was used to include (in 8 additional separate models) sex, dual status, sex + dual status, sex + dual status + race, sex + dual status + income + education + unemployment, sec + dual status + AHRQ SES Index, sex + dual status + income + education + unemployment + race, and sex + dual status + race + AHRQ SES Index. The developers report that the analyses found the relationship between the various social risk factors tested and the measure cost scores were inconsistent across factors and sometimes negative. They claim including these factors could introduce bias into the measure. There were many significant p values that indicate social risk factor are predictive of resource use. However, analysis results suggested that adding social risk factors to the measure risk-adjustment model had minimal impact on measure performance and was largely redundant with current model prediction. This was determined using two methods: by (i) analyzing differences in percentiles of observed to expected episode cost (O/E) ratios both with and without social risk factors in the model, and (ii) examining correlations between measure scores calculated with and without social factors. Both of these tests demonstrated a minimal impact on performance – even for providers at high and low extremes of risk - from including social risk factors in the model. Under the first test, the majority of providers – 95.0 percent of TINs and 94.5 percent of TIN-NPIs – saw no or minimal change (5 percentiles or less) in performance percentile when social risk factors were added to the model. Under the second test, measure scores calculated with and without social factors were highly correlated at both the TIN and TIN-NPI levels, with Spearman correlation coefficients of 0.996 and 0.996, respectively. DECISION: not include social risk factors in the model. Given the testing completed, this makes sense. However, the data used at Census Block level may not be precise enough to capture true relationships and there were many significant p values. I am not concerned about the different results for different social risk factors, this is common for various underlying reasons. The overall R-squared for the measure was 0.516 (and adjusted R-squared 0.513). The average observed to expected cost was generally close to one, 0.99 to 1.01, across risk deciles, indicating that the model is accurately predicting actual episode cost across risk deciles, and the average O/E cost ratios for all risk deciles are close to 1.0.

20. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

    Submission document: Testing attachment, section 2b4.

    Panel Member 1: None.

    Panel Member 2: None.

    Panel Member 4: No concerns

    Panel Member 5: None. There are clinically and practically significant variation in Lumbar Spine Fusion Measure scores, indicating the measure's ability to capture differences in performance. Our findings regarding variation in measure scores are consistent with expert clinician input and the face validity rating from expert clinicians that scores obtained from the measure specifications will provide an accurate reflection of the cost of episodes of care, and can be used to distinguish good and poor performance on cost effectiveness

    Panel Member 6: See #3623 for concerns.

    Panel Member 8: The authors do show sub-groups at the top and bottom of the distribution which appear to be different from the large group of TIN or TIN_NPIs in the middle -- this is a typical pattern for episode based resource use measures.

    Panel Member 9: The Lumbar Spine Fusion measure scores have a good deal of variability. For TINs, the standard deviation is 0.09, and 99/1, 90/10, and 75/25 percentile ratios are 1.56, 1.23, and 1.10, respectively. For TIN-NPIs, the standard deviation is 0.10, and 99/1, 90/10, and 75/25 percentile ratios are 1.56, 1.24, and 1.11, respectively. Scores were not influenced by region or number of episodes performed.

21. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

    Submission document: Testing attachment, section 2b5.

    Panel Member 1: None.

    Panel Member 2: NA

    Panel Member 4: N/A

    Panel Member 5: NA

    Panel Member 5: NA

    Panel Member 8: NA.

22. **Please describe any concerns you have regarding missing data.**

    Submission document: Testing attachment, section 2b6.

    Panel Member 1: None.

    Panel Member 2: NA

    Panel Member 4: No concerns

    Panel Member 5: None

    Panel Member 6: NA

    Panel Member 8: None.

For cost/resource use measures ONLY:

23. **Are the specifications in alignment with the stated measure intent?**

    ☒ **Yes** ☒ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

    Panel Member 6: See #3623 for attribution concerns

24. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

    Panel Member 1: Need discussion about the excluding the cost of a patient that dies within the care episode if such death is due to complications associated with the surgery.

    Panel Member 2: As noted above, we do not have adequate documentation of how costs are narrowly defined to only be those associated with Lumbar Fusion and its post-acute care. TEP vote on this issue, service assignment, saw greatest split 6-3

    Panel Member 5: None

    Panel Member 6: See #3623 for attribution concerns

    Panel Member 8: The measure includes price standardization, which is important. The attribution method focuses on the lead surgeon and co-surgeons, dropping any case where the surgeon is not clear. It is not clear why other clinicians on the surgical team do not also have the measure attributed to them.

25. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

    ☐ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

    ☒ **Low** (NOTE: Should rate LOW if you believe that there **are** threats to validity and/or relevant threats to validity were **not assessed OR** if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level **is required**; if not conducted, should rate as INSUFFICIENT.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

    Panel Member 1: Based on my assessments of validity sub criteria in #12 through #24.

    Panel Member 3: To demonstrate a moderate level, the developer must show an empirical association between the implicit quality construct and the material outcome

    Panel Member 4: Very modest correlation between measure and MSPB.

    Panel Member 5: Reasonable approaches for type of measure and newness

    Panel Member 6: See #3623 for attribution concerns

    Panel Member 8: This early evidence suggests moderate validity. Further evidence over time will help determine if this is a 'high validity' measure.

    Panel Member 9: The overall face validity rating of 4.5 indicates somewhat to moderate agreement that the measure as specified would provide an accurate reflection of costs for episodes of care and ability to distinguish good and poor performance on cost effectiveness. The empirical validity analysis showed very similar mean cost scores (observed to expected) for low performing to high performing TINs and TIN/NPIs, ranging from 1.04 to 0.96 from worst to best cost performance groups for TINs and 1.04 to 0.96 from worst to best cost performance for TIN/NPIs (total range of 8% difference).

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

    Panel Member 2: Per comments above, how costs are narrowed to those related to lumbar fusion. Social determinants in risk model

    [Response Ends]

Developer Submission

Brief Measure Information

[Response Begins]

NQF #:

De.2. **Measure Title:**

Co.1.1. **Measure Steward:**

De.3. **Brief Description of Measure:**

IM.1.1. **Developer Rationale:**

De.1. **Measure Type:**

S.5. **Data Source:**

S.3. **Level of Analysis:**

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?**

[Response Ends]

Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

IM.1. **Opportunity for Improvement**

IM.1.1. **Briefly explain the rationale for this measure (e.g., the benefits or improvements in performance envisioned by use of this measure)**

[Response Begins]

Lumbar spine fusion surgeries comprise some of the largest admission expenditures in the Medicare program, and are increasingly prevalent among Medicare patients. Total admission expenditures for these procedures exceeded $3.6 billion in 2013, and more than 6 million Medicare patients were diagnosed with lumbar degenerative conditions between 2006 and 2012. [1][2] Currently, there are substantial opportunities to improve the cost-efficiency and quality of care related to these procedures, given the high variation in treatment options. Primarily these include the use of less-invasive surgical techniques to reduce post-operative complications. [3]

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode-based cost measure (also referred to in this form as “the Lumbar Spine Fusion” measure) was recommended for development by an expert clinician committee (the Musculoskeletal Disease Management - Spine Clinical Subcommittee, composed of 22 clinician experts affiliated with 19 specialty societies) because of its impact in terms of Medicare cost and patient populations, as well as the opportunity for incentivizing cost-effective, high quality care in this area. Based on the initial recommendations from the Clinical Subcommittee, a subsequent Lumbar Spine Fusion clinician expert workgroup (composed of 13 members affiliated with 13 specialty societies) provided extensive, detailed input on this measure. Workgroup input has helped ensure the measure's ability to fairly evaluate clinician cost performance for Lumbar Spine Fusion surgeries and to promote efficient and high quality care for Medicare patients undergoing these procedures.

[1] Zorica Buser et al., "Spine Degenerative Conditions and Their Treatments: National Trends in the United States of America." [In eng]. Global Spine J 8, no. 1 (Feb 2018): 57-67.

[2] Steven D. Culler et al., "Incremental Hospital Cost and Length-of-Stay Associated with Treating Adverse Events among Medicare Beneficiaries Undergoing Lumbar Spinal Fusion During Fiscal Year 2013." [In eng]. Spine (Phila Pa 1976) 41, no. 20 (Oct 15 2016): 1613-20.

[3] Christina L. Goldstein et al., "Comparative Outcomes of Minimally Invasive Surgery for Posterior Lumbar Fusion," Clinical Orthopaedics and Related Research," 472, no.6 (2014): 1727-1737, https://doi.org/10.1007/s11999-014-3465-5.

IM.1.2. Provide performance scores on the measure as specified (current and over time) **at the specified level of analysis.** (This is required for endorsement maintenance. Include mean, stddev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include). **This information also will be used to address the subcriterion on improvement (U.3.1.) under Usability and Use.**

Performance scores are provided for clinician groups (identified by Tax Identification Number [TIN]) and individual clinicians (identified by a combination of TIN and National Provider Identifier [NPI]) attributed 10 or more Lumbar Spine Fusion episodes, as identified in Medicare Parts A and B claims data, ending from January 1, 2019, to December 31, 2019. These scores reflect 1,415 clinician group practices and 3,330 individual practitioners, corresponding to 54,768 episodes of care for 54,768 beneficiaries. Episodes are included from all 50 States and D.C. in the following settings: acute IP hospitals, OP facilities, ambulatory/office-based care centers, and ambulatory surgical centers (ASC).

TIN Level Scores:

- Mean score: 1.01
- Standard deviation: 0.09
- Min score: 0.62
- Max score: 1.54
- Score IQR: 0.10
- Score deciles:
    - 10th: 0.91
    - 20th: 0.94
    - 30th: 0.96
    - 40th: 0.98
    - 50th: 1.00
    - 60th: 1.02
    - 70th: 1.04
    - 80th: 1.07

- 90th: 1.12

TIN-NPI Level Scores:

- Mean score: 1.00
- Standard deviation: 0.10
- Min score: 0.62
- Max score: 1.84
- Score IQR: 0.11
- Score deciles:
  - 10th: 0.90
  - 20th: 0.93
  - 30th: 0.95
  - 40th: 0.97
  - 50th: 0.99
  - 60th: 1.01
  - 70th: 1.03
  - 80th: 1.06
  - 90th: 1.12

IM.1.3. **If no or limited performance data on the measure as specified is reported in IM.1.2., then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

N/A

IM.1.4. **Provide disparities data from the measure as specified** (current and over time) **by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** (This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) **This information also will be used to address the subcriterion on improvement (U.3.1.) under Usability and Use.**

N/A

IM.1.5. **If no or limited data on disparities from the measure as specified is reported in IM.1.4., then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

N/A

IM.2. Measure Intent

IM.2.1. Describe intent of the measure and its components/ Rationale (including any citations) for analyzing variation in resource use in this way.

The Lumbar Spine Fusion measure was developed for use in MIPS in the QPP to meet the requirements of the Social Security Act section 1848(r), added by the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). MIPS aims to reward high-value care by measuring clinician performance through four areas: quality, improvement activities, promoting interoperability, and cost. Each category assesses different aspects of care, and the categories are weighted such that they are combined into one composite score. CMS is introducing MIPS Value Pathways (MVPs) as a way to align and connect quality measures, cost measures, and improvement activities across performance categories of MIPS for different specialties or conditions. MVPs aim to provide a holistic assessment of clinician value for a specific type of care to achieve better healthcare outcomes and lower costs for patients. The use of cost measures is required by statute. The purpose of a cost measure as defined by NQF is to assess resource use. To be effective, they should capture costs related to a clinician's care decisions and account for factors outside of their influence.

Rationale for Measuring Cost through Episode-Based Cost Measures

The intent of an episode-based cost measure is to assess costs for a particular type of care, such as related to a procedure or the care of a condition. To do this, the measure only includes the cost of services that are clinically related to the role of the attributed clinician in providing care to a beneficiary. This is a key difference from broad, population-based cost measures such as the MIPS Total Per Capita Cost (TPCC) and Medicare Spending Per Beneficiary (MSPB) Clinician measures, which assess the overall costs of primary and inpatient care, respectively. Episode- and population-based measures complement each other, as they focus on different types of care.

Rationale for Measuring Cost of Lumbar Spine Fusion

Lumbar spine fusion surgeries are increasingly prevalent among Medicare patients and make up a large share of Medicare admission spending. This is an important area of cost to assess given the frequency of this procedure, the high costs associated with surgery, and the opportunities for clinicians to make care decisions that reduce the likelihood of high costs, as identified through expert stakeholder input and supported by the literature. Primary opportunities for improvement include the reduction of post-operative complications and readmissions.

This measure provides clinician with information about their costs of care that they can use to understand costs associated with their decision-making. Clinicians play an important role in variation in health care expenditures due to their ability to affect the costs associated with this surgery. [4] Between 2006 and 2012, over 6 million Medicare patients were diagnosed with lumbar degenerative conditions [5], and lumbar spine procedures are increasingly used in older adult patients to treat these conditions. One study found that 5.9 per 100 patients progressed to lumbar fusion within one year of diagnosis with lumbar degeneration, and there was an increase of 18.5% in the incidence of fusion procedures within one year of diagnosis [6]. Based on a review of the Medicare Provider Analysis and Review file, total spending on lumbar spinal fusion surgery is also one of the highest admission expenditures in the Medicare program, costing over $3.6 billion dollars in 2013 [7]. A systematic review comparing minimally invasive surgical (MIS) approaches to the lumbar spine for posterior fusion to open transforaminal or posterior lumbar interbody fusions and found that while surgical times and postoperative pain were similar between the two cohorts, MIS produced fewer complications and adverse medical events. [8]

Rationale for Use of Claims Data to Measure Cost

- The use of claims data for episode-based cost measures for MIPS is required by MACRA section 101(f).
- There is no additional submission burden, as clinicians must already submit claims for reimbursement.
- Using Medicare Parts A and B claims data allows CMS to evaluate TIN and TIN-NPI cost across all conditions and procedures, resulting in a comprehensive set of data on Lumbar Spine Fusion cost performance.

- Additionally, the wide reach of Medicare claims data maximizes the impact of the measure, ensuring that the most TINs and TIN-NPIs benefit from the information provided on Lumbar Spine Fusion cost performance.

[4] David Cutler et al., "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," American Economic Journal: Economic Policy 11, no. 1 (February 1, 2019): 192–221, https://doi.org/10.1257/pol.20150421.

[5] Buser et al., "Spine Degenerative Conditions and Their Treatments: National Trends in the United States of America," 57-67.

[6] Ibid.

[7] Culler et al., "Incremental Hospital Cost and Length-of-Stay Associated with Treating Adverse Events among Medicare Beneficiaries Undergoing Lumbar Spinal Fusion During Fiscal Year 2013," 1613-20.

[8] Goldstein et al., "Comparative Outcomes of Minimally Invasive Surgery for Posterior Lumbar Fusion," 1727-1737.

[Response Ends]

Scientific Acceptability of Measure Properties

Extent to which the measure, **as specified**, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

[Response Begins]

De.5. Subject/Topic Area *(check all the areas that apply):*

De.6. Non-Condition Specific *(check all the areas that apply):*

De.7. Care Setting *(Select all the settings for which the measure is specified and tested):*

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

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

On the QPP Resource Library https://qpp.cms.gov/resources/resource-library, refer to PY2021 and Cost category for the "2021 MIPS Cost Information Forms" and "2021 MIPS Cost Measure Code Lists" ZIP files. Open files ending in "-l-fusion".

S.2. Type of resource use measure *(Select the most relevant)*

Per episode

S.3. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED):*

Clinician: Group/Practice, Clinician: Individual

S.4. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

S.5. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.5.1.

Claims

S.5.1. Data Source or Collection Instrument *(Identify the specific data source or data collection instrument, e.g. name of database, clinical registry, collection instrument, etc.)*

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure uses Medicare Part A and Part B claims data, which is maintained by the Centers for Medicare & Medicaid Services (CMS). Part A and B claims data are used to build episodes of care, calculate episode costs, and construct risk adjustors. Data from the Medicare Enrollment Database (EDB) are used to determine patient-level exclusions and supplemental risk adjustors, specifically Medicare Parts A, B, and C enrollment; primary payer; disability status; end-stage renal disease (ESRD); patient birth dates; and patient death dates. The risk adjustment model also uses information from the Minimum Data Set (MDS) to account for expected differences in payment for services provided to patients in long-term care via a Long Term Care Indicator variable in risk adjustment.

For measure testing, data from the United States Census Bureau American Census, United States Census Bureau American Community Survey (ACS), and Common Medicare Enrollment (CME) are used in the analyses evaluating social risk factors in risk adjustment.

S.5.2. Data Source or Collection Instrument Reference *(available at measure-specific Web page URL identified in S.1 OR in the file attached here) (Save file as: S_5_2_DataSourceReference)*

<SamplingMethodologySpecificDataSourceAttachment nodeType="0">2020-12-09-codes-list-l-fusion.xlsx

S.6. Data Dictionary or Code Table *(Please provide a web page URL or attachment if exceeds 2 pages. NQF strongly prefers URLs. Attach documents only if they are not available on a web page.)*

Data Dictionary:

URL: The Research Data Assistance Center (ResDAC) maintains the Medicare claims data dictionary available here: https://www.resdac.org/file-availability-vrdc. CMS maintains the Medicare EDB and data dictionary: edbonline@cms.hhs.gov.

Please supply the username and password:

Attachment:

Code Table:

URL:

Please supply the username and password:

Attachment: 2021-01-08-testing-form-appendix-l-fusion.xlsx

Construction Logic

S.7.1. Brief Description of Construction Logic

If applicable, summarize the general approach or methodology to the measure construction. This is most relevant to measures that are part of or rely on the execution of a measure system or applies to multiple measures.

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure is the risk-adjusted cost across all episodes attributed to the clinician group (identified by Taxpayer Identification Number, or TIN) or individual clinician (identified by unique combination of Taxpayer Identification Number and National Provider Identifier, or TIN-NPI).

Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes, which are units or specific instance of the measure for a given patient and clinician or clinician group are triggered or opened by Current Procedural Terminology / Healthcare Common Procedure Coding System (CPT/HCPCS) codes indicating the presence of a lumbar spine fusion procedure. The episode window spans from 30 days prior to the trigger day through 90 days after, and includes costs from certain clinically-related services from Medicare Parts A and B claims during the episode window.[1] Cost figures are standardized to account for differences in Medicare payments for the same service(s) across Medicare providers. Payment standardized costs remove the effect of differences in Medicare payment among health care providers that are the result of differences in regional health care provider expenses measured by hospital wage indexes and geographic price cost indexes (GPCIs) or other payment adjustments such as those for teaching hospitals. This standardization is intended to isolate cost differences that result from healthcare delivery choices, allowing for more accurate resource use comparisons between health care providers. [2] A regression model is applied to estimate the expected cost of each episode for risk adjustment.

The cost measure is calculated as the sum of the ratio of observed to expected payment-standardized cost to Medicare for all Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes attributed to a clinician or clinician group. The resulting average episode cost ratio is then multiplied by the national average observed episode cost to generate a dollar figure.

[1] Cost is defined by allowed amounts on Medicare claims data, which include both Medicare trust fund payments and any applicable beneficiary deductible and coinsurance amounts. Claims data from Medicare Parts A and B are used to construct the episode-based cost measures.

[2] For more information on payment standardized costs, please refer to the “CMS Price (Payment) Standardization - Basics" and “CMS Price (Payment) Standardization - Detailed Methods” documents posted on the CMS Price (Payment) Standardization Overview page (https://www.resdac.org/articles/cms-price-payment-standardization-overview).

S.7.2. Construction Logic *(Detail logic steps used to cluster, group or assign claims beyond those associated with the measure’s clinical logic.)*

Step 1. Trigger and Define an Episode

Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes are defined by CPT/HCPCS codes on Part B Physician/Supplier (Carrier) claims that open, or trigger, an episode.

The steps for defining an episode for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode group are as follows:

- Identify Part B Physician/Supplier claim lines with positive standardized payment that have a trigger code.
- Trigger an episode if all the following conditions are met for an identified Part B Physician/Supplier claim line:
    - It was billed by a clinician of a specialty that is eligible for the Merit-based Incentive Payment System (MIPS).
    - It does not have a post-operative modifier code. [3]
    - It is the highest cost claim line across all claim lines identified in the above bullets and that have any Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels trigger code billed for

the patient on that day. If multiple Part B Physician/Supplier claim lines with a trigger code occur on different days within a concurrent inpatient (IP) stay, an episode will be triggered by the claim line with the earliest expense date during the IP stay.

- Identify episodes that have a concurrent IP stay by identifying the first IP stay with a relevant Medicare Severity Diagnosis-Related Group (MS-DRG) code for the patient that is concurrent to the expense date for the trigger Part B Physician/Supplier claim line.
- Establish the episode window as follows:
  - Establish the episode trigger date as the IP start day if an IP stay with a relevant MS-DRG concurrent with the trigger is found, otherwise the expense date of the trigger code.
  - Establish the episode start date as 30 days prior to the episode trigger date.
  - Establish the episode end date as 90 days after the episode trigger date.

Once a Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode is triggered, the episode is placed into one of the episode sub-groups to enable meaningful clinical comparisons. Sub-groups represent more granular, mutually exclusive and exhaustive patient populations defined by clinical criteria (e.g., information available on the patient's claims at the time of the trigger). Sub-groups are useful in ensuring clinical comparability so that the corresponding cost measure fairly compares clinicians with a similar patient case-mix. This cost measure has 3 sub-groups:

- Level 1 Lumbar Fusion
- Level 2 Lumbar Fusion
- Level 3 Lumbar Fusion

To provide further background on these sub-group classifications, the lumbar region of the spine generally consists of five lumbar vertebrae. A single level (Level 1) procedure refers to the fusion of one segment of the spine to join two vertebrae (e.g., L5-S1). A Level 2 procedure refers to the fusion of two segments of the spine (e.g., L4-L5 and L5-S1), and a Level 3 procedure refers to the fusion of three segments of the spine (e.g., L3-L4, L4-L5, L5-S1). In claims data, Level 1 procedures are identified by the presence of a CPT/HCPCS code. Procedures in Level 2 and Level 3 are identified by the presence of a CPT/HCPCS code plus add-on codes to account for multiple services.

Step 2. Attribute Episodes to a Clinician

Once an episode has been triggered and defined, it is attributed to one or more clinicians of a specialty that is eligible for MIPS. Clinicians are identified by TIN-NPI, and clinician groups are identified by TIN. Only clinicians of a specialty that is eligible for MIPS or clinician groups where the triggering clinician is of a specialty that is eligible for MIPS are attributed episodes.

The steps for attributing a Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode are as follows:

- Identify claim lines with positive standardized payment for any trigger codes that occur during the IP stay, if the triggering procedure occurs during an IP stay with a relevant MS-DRG, otherwise identify claim lines with positive standardized payment for any trigger codes that occur on the trigger day.
- Designate a TIN-NPI as a main clinician if the following conditions are met:
  - No assistant modifier code is found on one or more claim lines billed by the clinician.
  - No exclusion modifier code is found on the same claim line.
- Designate a TIN-NPI as an assistant clinician if the following conditions are met:

  - The TIN-NPI was not designated as a main clinician.
  - An assistant modifier code is found.
  - No exclusion modifier code is found.
- Attribute an episode to any TIN-NPI designated as a main or assistant clinician.
- Attribute episodes to the TIN by aggregating all episodes attributed to NPIs that bill to that TIN. If the same episode is attributed to more than one NPI within a TIN, the episode is attributed only once to that TIN.

Step 3. Assign Costs to an Episode and Calculate Total Observed Episode Cost

Services, and their Medicare costs, are assigned to an episode only when clinically related to the attributed clinician's role in managing patient care during the episode. Assigned services may include treatment and diagnostic services, ancillary items, services directly related to treatment, and those furnished as a consequence of care (e.g., complications, readmissions, unplanned care, and emergency department visits). Unrelated services are not assigned to the episode. For example, the cost of care for a chronic condition that occurs during the episode but is not related to the clinical management of the patient relative to the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels would not be assigned.

For the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode group, only services performed in the following service categories are considered for assignment to the episode costs:

- Emergency Department (ED)
- Outpatient (OP) Facility and Clinician Services
- IP - Medical
- IP - Surgical
- Inpatient Rehabilitation Facility (IRF) - Medical
- Durable Medical Equipment, Prosthetics, Orthotics, and Supplies (DME)
- Home Health (HH)

In addition to service category, service assignment rules may be modified based on the service category in which the service is performed, as listed above. Service assignment rules may also be defined based on specific (i) service information alone or service information combined with diagnosis information, (ii) prior incidence of service, and/or (iii) the timing of the service, as detailed below.

- Services may be assigned to the episode based on the following service information combinations:
  - High level service code alone
  - High level service code combined with first 3 digits of the International Classification of Diseases – 10th Revision diagnosis code (3-digit ICD-10 diagnosis code)
  - High level service code combined with full ICD-10 diagnosis code
  - High level service code combined with more specific service code
  - High level service code combined with more specific service code and with 3-digit ICD-10 diagnosis code
  - High level service code combined with more specific service code and with full ICD-10 diagnosis code
- Assigned services may be further refined by prior incidence of service or diagnosis:

    - Services may be assigned unconditionally (regardless of prior incidence of the service in patient's recent claims history).
    - Services may be assigned if newly occurring.
    - Services may be assigned in combination with a diagnosis if the service is newly occurring.
    - Services may be assigned in combination with a diagnosis if the diagnosis is newly occurring.
    - Services may be assigned in combination with a diagnosis if either the service OR the diagnosis are newly occurring.
    - Services may be assigned in combination with a diagnosis if both the service AND the diagnosis are newly occurring.
- Services as defined by the applicable combinations and incidence options above may be assigned with only specific timing:
    - Services may be assigned based on whether or not the service occurs before the trigger (in the pre-trigger window) and/or after the trigger (in the post-trigger window).
    - Services may be assigned only if they occur within a particular number of days from the trigger within the episode window, and services may be assigned for a period shorter than the full duration of the episode window.
    - The full list of service assignment rules for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure can be found on the "Service_Assignment" tab of the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure Codes List file.
    - The steps for assigning costs are as follows:
- Identify all services on claims with positive standardized payment that occur within the episode window.
- Assign identified services to the episode based on the types of service assignment rules described above.
- Assign skilled nursing facility (SNF) claims based on the following criteria:
    - Identify SNF claims for which both (i) the SNF claim's qualifying IP stay is the IP stay during which the trigger occurs, if an IP stay is found, and (ii) the SNF claim occurs during the episode window.
    - For those identified SNF claims, assign the percentage of the claim amount proportional to the portion of the SNF claim that overlaps with the episode window.
- Assign all claims with trigger codes occurring during the trigger day/stay.
- Assign all physician claims and DME claims occurring during concurrent IP stay as applicable.
- Assign all IP evaluation and management (E&M) claims during IP stays in the post-trigger window assigned to episode.
- Sum standardized Medicare allowed amounts for all claims assigned to each episode to obtain the standardized total observed episode cost.

Step 4. Exclude Episodes

Before measure calculation, episode exclusions are applied to remove certain episodes from measure score calculation. Certain exclusions are applied across all procedural episode groups, and other exclusions are specific to the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels, based on consideration of the clinical characteristics of a homogenous patient cohort. The measure-specific exclusions are listed in the "Exclusions" and "Exclusions_Details" tabs in the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure Codes List file (referenced in Section S.1).

The steps for episode exclusion are as follows:

- Exclude episodes from measure calculation if:
    - The patient has a primary payer other than Medicare for any time overlapping the episode window or 120-day lookback period prior to the trigger day.
    - The patient was not enrolled in Medicare Parts A and B for the entirety of the lookback period plus episode window, or was enrolled in Part C for any part of the lookback plus episode window.
    - No main clinician is attributed the episode.
    - The patient's date of birth is missing.
    - The patient's death date occurred before the episode ended.
    - The episode trigger claim was not performed in an ambulatory/office-based care, IP hospital, OP hospital, or ASC setting based on its place of service.
    - The IP facility is not a short-term stay acute hospital as defined by subsection (d) when an IP stay concurrent with the trigger is found [4].
- Apply measure-specific exclusions, which check the patient's Medicare claims history for certain billing codes (as specified in the Measure Codes List file) that indicate the presence of a particular procedure, condition, or characteristic.

Step 5. Estimate Expected Costs through Risk Adjustment

Risk adjustment is used to estimate expected episode costs in recognition of the different levels of care patients may require due to comorbidities, disability, age, and other risk factors. The risk adjustment model includes variables from the CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) 2016 Risk Adjustment Model [5], as well as other standard risk adjustors (e.g., patient age) and variables for clinical factors that may be outside the attributed clinician´s reasonable influence. A full list of risk adjustment variables can be found in the "RA" and "RA_Details" tabs of the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure Codes List file (referenced in Section S.1).

Steps for defining risk adjustment variables and estimating the risk adjustment model are as follows:

- Define HCC and episode group-specific risk adjustors using service and diagnosis information found on the patient's Medicare claims history in the 120-day period prior to the episode trigger day for certain billing codes that indicate the presence of a procedure, condition, or characteristic.
- Define other risk adjustors that rely upon Medicare beneficiary enrollment and assessment data as follows:
    - Identify patients who are originally "Disabled without end-stage renal disease (ESRD)" or "Disabled with ESRD" using the original reason for joining Medicare field in the Medicare beneficiary EDB.
    - Identify patients with ESRD if their enrollment indicates ESRD coverage, ESRD dialysis, or kidney transplant in the Medicare beneficiary EDB in the lookback period.

    - Identify patients who have spent at least 90 days in a long-term care institution without having been discharged to the community for 14 days, based on LTC MDS assessment data, during the lookback period.
- Drop risk adjustors that are defined for less than 15 episodes nationally to avoid using very small samples.
- Categorize patients into age ranges using their date of birth information in the Medicare beneficiary EDB. If an age range has a cell count less than 15, collapse this with the next adjacent higher age range category towards the reference category (65-69).
- Include the MS-DRG of the episode's trigger IP stay, if an IP stay is found, as a categorical risk adjustor.
- Run an ordinary least squares (OLS) regression model to estimate the relationship between all the risk adjustment variables and the dependent variable, the standardized observed episode cost, to obtain the risk-adjusted expected episode cost. A separate OLS regression is run for each episode sub-group nationally.
- Winsorize [6] expected costs as follows.
    - Assign the value of the 0.5th percentile to all expected episode costs below the 0.5th percentile.
    - Renormalize [7] values by multiplying each episode´s winsorized expected cost by the average expected cost, and dividing the resultant value by the average winsorized expected cost.
- Exclude [8] episodes with outliers as follows. This step is performed separately for each sub-group.
    - Calculate each episode´s residual as the difference between the re-normalized, winsorized expected cost computed above and the observed cost.
    - Exclude episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution.
    - Renormalize the resultant expected cost values by multiplying each episode's winsorized expected costs after excluding outliers by the sub-group's average standardized observed cost after excluding outliers, and dividing by the sub-group's average winsorized expected cost after excluding outliers.

Step 6. Calculate Measure Scores

Measure scores are calculated for a TIN or TIN-NPI as follows:

- Calculate the ratio of observed to expected episode cost for each episode attributed to the clinician/clinician group.
- Calculate the average ratio of observed to expected episode cost across the total number of episodes attributed to the clinician/clinician group.
- Multiply the average ratio of observed to expected episode cost by the national average observed episode cost to generate a dollar figure representing risk-adjusted average episode cost.

[3] Post-operative modifier codes indicate that a clinician billing the service was not involved in the main procedure but was involved in the post-operative care for that procedure, and as such the post-operative clinician would not be responsible for the trigger.

[4] Only stays at IP facilities that are paid under a short-term stay acute hospital as defined by subsection (d) will be included. Subsection (d) hospitals are hospitals in the 50 states and D.C. other than: psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer. For details on the identification of these hospitals, please refer to the CMS Certification Number (CCN) definitions for Short-term (General and Specialty) Hospitals facility types in Section 2779A1 of Chapter 2 of the CMS State Operation Manual. (https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Downloads/som107c02.pdf).

[5] CMS uses an HCC risk adjustment model to calculate risk scores. The HCC model ranks diagnoses into categories that represent conditions with similar cost patterns. Higher categories represent higher predicted healthcare costs, resulting in higher risk scores. There are over 9,500 ICD-10-CM codes that map to one or more of the 79 HCC codes included in the CMS-HCC V22 model.

[6] Winsorization aims to limit the effects of extreme values on expected costs. Winsorization is a statistical transformation that limits extreme values in data to reduce the effect of possible outliers. Winsorization of the lower end of the distribution (i.e., bottom coding) involves setting extremely low predicted values below a predetermined limit to be equal to that predetermined limit.

[7] Renormalization is performed after adjustments are made to the episode's expected cost, such as bottom-coding or residual outlier exclusion. This process multiplies the adjusted values by a scalar ratio to ensure that the resulting average is equal to the average of the original value.

[8] This step excludes episodes based on outlier residual values from the calculation and renormalizes the resultant values to maintain a consistent average episode cost level.

S.7.2a. CONSTRUCTION LOGIC ATTACHMENT or URL: If needed, attach supplemental documentation (Save file as: S_7_2_Construction_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.

URL:

Please supply the username and password:

Attachment:

S.7.3. Concurrency of clinical events, measure redundancy or overlap, disease interactions *(Detail the method used for identifying concurrent clinical events, how to manage them, and provide the rationale for this methodology.)*

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels includes costs from clinically related Medicare Part A and Part B services that are furnished to a patient during the episode. The measure avoids redundancy or overlap of clinical events by counting each service once within a given episode for the attributed clinician(s).

This measure is designed to allow episodes to overlap with other episodes; overlapping episodes are different episodes that are triggered for the same patient with overlapping episode windows. This approach allows each episode to reflect attributed clinicians' different roles in providing care services throughout a patient's care trajectory and ensures continuous accountability throughout a patient's care. For example, a patient could have an Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode triggered when the attributed clinician performs the procedure, and 80 days later be admitted to hospital for pneumonia unrelated to the lumbar spine fusion, triggering an episode for a different cost measure that is attributed to the hospitalist providing care for pneumonia. Each episode includes only the cost of assigned services (i.e., those that are within the reasonable influence of the attributed clinician) to reflect each attributed clinician's role. In addition, costs are not double counted as the measure calculation is based on the ratio of observed over expected spending for each episode, then averaged across all of an attributed clinician's episodes.

The measure also allows for multiple procedure types to occur, such as same-day anterior and posterior lumbar fusions, and combined approaches, which are accounted for in risk adjustment due to increased complexity.

The measure accounts for disease interactions through its risk adjustment model based on the CMS-HCC V22 2016 model. In addition to the HCCs, the model includes disease interactions (e.g., Cancer * Immune Disorders). Further details about the risk adjustment model and disease interaction terms are included in Section S.8.6.

S.7.4. Complementary services *(Detail how complementary services have been linked to the measure and provide rationale for this methodology.)*

This measure includes the cost of services that are clinically related to the procedure for Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels. The rationale for only including specific costs is to ensure that the attributed clinician is evaluated only on his or her performance on services over which they have reasonable influence, or can reasonably influence the frequency or severity. For instance, the cost of an emergency visit for post-operative infection is included in a clinician's episode cost if it occurs within 7 days of the procedure.

These assigned services that have been identified as related to the procedure and within the influence of the attributed clinician were identified based on empirical evidence and detailed clinical input, the latter of which was gathered from clinician experts and broader feedback from stakeholders from the clinician community. The list of assigned services can be found in the "Service_Assignment" tab of the Measure Codes List linked in Section S.1, the construction logic used to calculate costs of assigned services is described in Step 3 of Section S.7.2, and the stakeholder input processes used to identify and refine these included services is described in Section S.8.3.

S.7.5. Clinical hierarchies *(Detail the hierarchy of codes or condition groups used and provide rationale for this methodology.)*

The risk adjustment model for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure includes variables from the CMS-HCC V22 2016 Risk Adjustment Model, as well as other standard risk adjustors (e.g., patient age brackets using information in the Medicare beneficiary EDB) and disease interaction terms. The model also includes variables specific to lumbar spine fusion, identified through the incorporation of detailed clinical input, for clinical conditions which may influence the procedure complexity, episode cost, and risk of complication. This approach is adopted to ensure sufficient capture of the patient's clinical characteristics prior to the episode and to allow more comprehensive risk adjustment of comorbid factors, such that remaining variation in clinicians' costs to Medicare are limited to costs that clinicians can reasonably influence. Additional information about the risk adjustment model is included in Section S.8.6.

S.7.6. Missing Data *(Detail steps associated with missing data and provide rationale for this methodology (e.g., any statistical techniques to impute missing data)*

Since CMS uses Medicare claims data to calculate the Lumbar Spine Fusion for Degenerative Disease, 1-3 cost measure, we expect a high degree of data completeness.

The data fields used to calculate measure (e.g., payment amounts, diagnosis and procedure codes, etc.) are included in all Medicare claims because clinicians only receive payments for complete claims.

Additional information regarding the method of testing to identify missing data is available in the Testing Form in Section 2b6.

CMS has in place several auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and to recoup any overpayments. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in this measure, including diagnosis and procedure codes and other elements that are consequential to payment. Specifically, CMS works with Zone Program Integrity Contractors (ZPICs), and formerly Program Safeguard Contractors (PSCs), to ensure program integrity; the agency also uses Recovery Audit Contractors (RACs) to identify and correct for underpayments and overpayments.

CMS also uses the Comprehensive Error Rate Testing (CERT) Program to ensure that Medicare payments are correct in accordance with coverage, coding, and billing rules. Between 2005 and 2020, CERT estimates that proper payment, which is payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments each year. The FY 2020 Medicare FFS program proper payment rate (based on data from July 2018-June 2019) was 93.7 percent. [9] CMS continues to perform successful corrective actions and give providers additional education to ensure accurate billing.

To further ensure the completeness and accuracy of data for each beneficiary who opens an episode, the measure excludes episodes where beneficiary date of birth information (an input to the risk adjustment model) cannot be found in the EDB or the beneficiary death date occurs before the episode trigger date (an indication of errant data).

The Lumbar Spine Fusion for Degenerative Disease, 1-3 measure also excludes episodes where the patient is enrolled in Medicare Part C or has a primary payer other than Medicare in the 120-day lookback period and episode window. In such situations, Medicare Parts A and B claims data may not capture the complete clinical profile for the patient needed to capture the clinical risk of the patient in risk adjustment. Furthermore, Parts A and B claims data may not capture all Medicare resource use if some portion of the patient's care is covered under Medicare Part C. These steps ensure that we have complete claims data for patients included in the Lumbar Spine Fusion for Degenerative Disease, 1-3 measure.

To ensure claims completeness and inclusion of any corrections, the measure was developed and calculated using data with a three month claims run-out from the end of the performance period.

[9] Comprehensive Error Rate Testing (CERT) Program. “2020 Medicare Fee-for-Service Supplemental Improper Payments Data”. Table A6. https://www.cms.gov/files/document/2020-medicare-fee-service-supplemental-improper-payment-data.pdf.

S.7.7. Resource Use Service Categories (Units) (Select all categories that apply)

- Inpatient services: Inpatient facility services
- Inpatient services: Evaluation and management
- Inpatient services: Procedures and surgeries
- Inpatient services: Imaging and diagnostic
- Inpatient services: Lab services
- Inpatient services: Admissions/discharges
- Other inpatient services
- Ambulatory services: Outpatient facility services
- Ambulatory services: Emergency Department

- Ambulatory services: Pharmacy
- Ambulatory services: Evaluation and management
- Ambulatory services: Procedures and surgeries
- Ambulatory services: Imaging and diagnostic
- Ambulatory services: Lab services
- Other ambulatory services
- Durable Medical Equipment (DME)
- Other services not listed
- See Measure Codes List
- See Measure Codes List
- See Measure Codes List

S.7.8. Identification of Resource Use Service Categories (Units)

(For each of the resource use service categories selected above, provide the rationale for their selection and detail the method or algorithms to identify resource units, including codes, logic and definitions.)

The Lumbar Spine Fusion for Degenerative Disease, 1-3 measure assesses the standardized allowed amounts of services by clinicians during an episode. Services are assigned (and their costs are included in the measure) only when clinically related to the attributed clinician's role in managing patient care during the episode from 30 days prior to the trigger day through 90 days after. The detailed logic conditions (service assignment rules) are included in the "Service_Assignment" tab of the Measure Codes list file (linked in Section S.1). This identification approach allows the measure to capture the cost of services that can be attributed to the clinician responsible for managing the patient's care before, during, and after the lumbar spine fusion, without capturing the cost of services that are considered clinically unrelated.

S.7.8a. If needed, provide supplemental resource use service category specifications in either URL (preferred) or as an attachment (Save file as S.7.8a_RU_Service_Categories):

URL: See URL provided in S.1.

Please supply the username and password:

Attachment:

Clinical Logic

S.8.1. Brief Description of Clinical Logic (Briefly describe your clinical logic approach including clinical topic area, whether or not your account for comorbid and interactions, clinical hierarchies, clinical severity levels and concurrency of clinical events.)

This measure aims to provide actionable information to clinicians performing a Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels about their resource use within the overall goal of enabling clinicians to provide cost-effective and high-quality care. The clinical logic is constructed to achieve this objective.

Clinical Topic Area: Lumbar spine fusion (1-3 levels)

Comorbidity and Interactions: The risk adjustment model includes a series of interaction terms between comorbidities and applies a variant of the CMS-HCC V22 risk adjustment model with additional risk adjustors specific to this procedure to capture patient comorbidities.

Clinical Hierarchies: Clinical hierarchies are embedded in the risk adjustment model, based on the CMS-HCC model.

Clinical Severity Levels: This measure has sub-groups to adjust for the levels of severity among Level 1, Level 2, and Level 3 lumbar fusions. A single level (Level 1) procedure refers to the fusion of one segment of the spine to join two vertebrae (e.g., L5-S1). A Level 2 procedure refers to the fusion of two segments of the spine (e.g., L4-L5 and L5-S1), and a Level 3 procedure refers to the fusion of three segments of the spine (e.g., L3-L4, L4-L5, L5-S1). It also risk adjusts for the MS-DRG when the procedure occurs in an inpatient setting, accounting for medical severity levels.

Clinical logic for the Lumbar Spine Fusion for Degenerative Disease, 1-3 measure counts each service once within a given episode for the attributed clinician(s). The measure also only includes services that are clinically related to the procedure defined by service assignment rules, which were specified based on input from the Lumbar Spine Fusion for Degenerative Disease, 1-3 Clinician Expert Workgroup.

S.8.2. Clinical Logic *(Detail any clustering and the assignment of codes, including the grouping methodology, the assignment algorithm, and relevant codes for these methodologies.)*

A spine fusion is a procedure that permanently fuses one or more vertebrae to stabilize the spine, reduce pain, and prevent nerve damage. While the lumbar region of the spine generally consists of five lumbar vertebrae, the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure focuses on Level 1, Level 2, or Level 3 spine fusion procedures. A single level (Level 1) procedure refers to the fusion of one segment of the spine to join two vertebrae (e.g., L5-S1). A Level 2 procedure refers to the fusion of two segments of the spine (e.g., L4-L5 and L5-S1), and a Level 3 procedure refers to the fusion of three segments of the spine (e.g., L3-L4, L4-L5, L5-S1). In claims data, Level 1 procedures are identified by the presence of a CPT/HCPCS code. Procedures in Level 2 and Level 3 are identified by the presence of a CPT/HCPCS code plus add-on codes to account for multiple services. This measure scope, along with exclusions described in S.9.1, ensures the measure focuses on procedures that are most common for degenerative disease and maintains a more homogenous patient cohort by excluding procedures performed for more complex patients.

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure evaluates resource use through the unit of episodes of care. The cost measure episodes are constructed by including the cost of assigned services provided by clinicians and other providers during the episode window, defined as 30 days prior to the episode trigger and 90 days after the trigger. Triggered episodes are attributed to one or more clinicians of a specialty that is eligible for MIPS, where individual clinicians are identified by TIN-NPI and clinician groups are identified by TIN, and the attributed clinician/clinician group renders the trigger CPT/HCPCS services. Within the specified episode window, the costs of clinically related pre-operative and follow-up services, including those that result as a consequence of care, such as post-surgical complications, would be assigned to the attributed clinician or clinician group using a service assignment algorithm. The episode triggers and assigned services are contained in the Measure Codes List file (see Section S.1. for details), along with codes used to aid in attribution, codes used to identify measure-specific risk adjustors (described in Section S.8.6), and codes used to identify exclusions (described in Section S.9.1).

The cost measure is calculated as the sum of the ratios of observed to expected costs, multiplied by the national average observed episode cost to generate a dollar figure, and then divided by total number of episodes from the episode group attributed to a clinician. All costs are payment standardized to control for geographic variation in Medicare reimbursement rates. The measure is risk adjusted to account for age and severity of illness. Expected costs are estimated through risk adjustment by using a linear regression model. More details about the risk adjustment model are described in Section S.7.5 and S.8.6.

S.8.3. Evidence to Support Clinical Logic Described in S.8.2 *Describe the rationale, citing evidence to support the grouping of clinical conditions in the measurement population(s) and the intent of the measure (as described in IM3)*

The clinical logic used in the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure is informed by literature, empirical data, expert input, and feedback from a range of stakeholders.

Cost measures are intended to help inform clinicians on the costs associated with their decision-making and to incentivize cost-effective, high-quality care. A cost measure offers opportunity for improvement if clinicians can exercise influence on the intensity or frequency of a significant share of costs during the episode, or if clinicians can achieve lower spending and better care quality through changes in clinical practice.

The measure was designed to incorporate extensive expert clinician input into each component of the measure to ensure that it achieves the goal of providing actionable information to clinicians for their performance of a procedure on a cohesive patient cohort. The measure was developed to meet the requirements of MACRA Section 101(f) to create episode-based cost measures. It aligns with CMS meaningful measure area of “patient-focused episode of care” within the overall quality priority of “Make Care Affordable.” The measure includes services that are clinically related to the procedure and within the reasonable influence of the attributed clinician. By including services after the procedure, it aims to improve care coordination throughout a patient’s care trajectory.

Between 2006 and 2012, over 6 million Medicare patients were diagnosed with lumbar degenerative conditions [9], and lumbar spine procedures are increasingly used in older adult patients to treat these conditions. One study found that 5.9 per 100 patients progressed to lumbar fusion within one year of diagnosis with lumbar degeneration, and there was an increase of 18.5% in the incidence of fusion procedures within one year of diagnosis [10]. Based on a review of the Medicare Provider Analysis and Review file, total spending on lumbar spinal fusion surgery is also one of the highest admission expenditures in the Medicare program, costing over $3.6 billion dollars in 2013 [11].

The Musculoskeletal Disease Management – Spine Clinical Subcommittee expert clinician committee, composed of 22 clinician experts affiliated with 19 specialty societies, recommended the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels for development because of its impact in terms of patient population and clinician coverage, and the opportunity for incentivizing cost-effective, high-quality clinical care in this area. Based on the initial recommendations from the Clinical Subcommittee, the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels clinician expert workgroup composed of 13 members affiliated with 13 specialty societies provided extensive, detailed input on this measure.

The members reviewed analyses of the utilization and timing of all Medicare Parts A and B services in broad timeframes extending before and after the episode trigger to provide input which services should be included as part of the episode costs. Members also provided clinical input on the particular logic conditions or rules that should be used along with the services, such as requiring additional codes to be present along with the service to ensure clinical relevance, assigning costs for the service if it occurs within a shorter timeframe from the trigger than the overall episode window length, or assigning the service only when accompanied by a particular relevant diagnosis that is newly occurring. Members also reviewed data on frequency and costs associated with sub-populations within the episode group’s patient cohort to inform input on risk adjustors and exclusions.

The draft measure was field tested from October to November 2018 along with several other measures; during this time, stakeholders reviewed the measure specifications, including a list of assigned services and associated logic rules, field test reports containing details of attributed clinician performance, and supplemental documentation. Over 75,000 TIN and TIN-NPI field test reports were available during this time for review and feedback.

During field testing, a National Summary Data Report, later updated to include reliability analyses, was posted along with the measure specifications:

- National Summary Data Report (2018) – this document contains summary data about Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure, along with other episode-based cost measures. These summary statistics supplement the testing analyses contained in this submission: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2018-national-summary-data-report.zip filename: 2020-06-05-national-summary-data-report.pdf.

  Stakeholder feedback gathered during field testing was summarized into the Field Testing Feedback Summary Report:

- Field Testing Feedback Summary Report (2018) – this document summarizes the feedback received during a stakeholder feedback period during measure development. The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure has been developed with extensive input from the clinician community: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2019-ft-feedback-summary-report.pdf.

Feedback gathered during field testing was evaluated by the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels clinician expert workgroup and considered in final measure specification refinements.

[9] Buser, Z., B. Ortega, A. D´Oro, W. Pannell, J. R. Cohen, J. Wang, R. Golish, M. Reed, and J. C. Wang. "Spine Degenerative Conditions and Their Treatments: National Trends in the United States of America." [In eng]. Global Spine J 8, no. 1 (Feb 2018): 57-67.

[10] Ibid.

[11] Culler, S. D., D. S. Jevsevar, K. G. Shea, K. J. McGuire, M. Schlosser, K. K. Wright, and A. W. Simon. "Incremental Hospital Cost and Length-of-Stay Associated with Treating Adverse Events among Medicare Beneficiaries Undergoing Lumbar Spinal Fusion During Fiscal Year 2013." [In eng]. Spine (Phila Pa 1976) 41, no. 20 (Oct 15 2016): 1613-20.

S.8.3a. CLINICAL LOGIC ATTACHMENT or URL: If needed, attach supplemental documentation (Save file as: S_8_3a_Clinical_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.

URL: See URL provided in S.1.

Please supply the username and password:

Attachment:

S.8.4. Measure Trigger and End mechanisms *(Detail the measure's trigger and end mechanisms and provide rationale for this methodology)*

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode is defined as follows:

- Episode trigger: CPT/HCPCS procedure code for a lumbar spine fusion (and if the procedure occurs during an IP stay, the admission must be relevant to the procedure as determined by a relevant MS-DRG code for spinal fusions (MS-DRGs 543-455, 459-460).

- Episode trigger date: IP admission date if an IP stay with a relevant DRG concurrent with the trigger is found, otherwise the expense date of the trigger code.
- Episode start date: 30 days prior to episode trigger date.
- Episode end date: 90 days after episode trigger date.

Additional conditions must be met to trigger an episode. Since the lumbar spine fusion procedure can occur in the inpatient or outpatient setting, a Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure can be triggered in the following settings: acute inpatient (IP) hospitals, hospital outpatient departments (HOPD), ambulatory/office-based care centers, and ambulatory surgical centers (ASC).

The detailed steps for triggering Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes are in Section S.7.2. The static timing of the episode start and end date are straightforward to ensure that clinicians can easily understand the episode window and construction, which is important for the goal of the measure to provide actionable information to clinicians.

The conditions to trigger episodes and the duration of the episode window were established with input from clinician experts in consideration of the goals of the measure to provide actionable information to clinicians about their resource use for a comparable patient cohort. An initial Draft List of Episode Groups and Trigger Codes was posted in December 2016 incorporating input from a Clinical Committee of more than 70 clinicians from over 50 professional societies. Feedback from a four-month public comment period on that posting was summarized and shared with clinical experts who used the information from the draft list as a starting point and took feedback into consideration along with analyses to help inform discussions (e.g., frequency of services over a period of time extending from the trigger date). This measure was field tested in 2018, as discussed further in Section S.8.3. Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels clinician expert workgroup took field testing feedback into consideration in making refinements to the measure, including feedback on episode exclusions and risk adjustors.The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode is defined as follows:

- Episode trigger: CPT/HCPCS procedure code for a lumbar spine fusion (and if the procedure occurs during an IP stay, the admission must be relevant to the procedure as determined by a relevant MS-DRG code for spinal fusions (MS-DRGs 543-455, 459-460).

- Episode trigger date: IP admission date if an IP stay with a relevant DRG concurrent with the trigger is found, otherwise the expense date of the trigger code.
- Episode start date: 30 days prior to episode trigger date.
- Episode end date: 90 days after episode trigger date.

Additional conditions must be met to trigger an episode. Since the lumbar spine fusion procedure can occur in the inpatient or outpatient setting, a Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure can be triggered in the following settings: acute inpatient (IP) hospitals, hospital outpatient departments (HOPD), ambulatory/office-based care centers, and ambulatory surgical centers (ASC).

The detailed steps for triggering Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes are in Section S.7.2. The static timing of the episode start and end date are straightforward to ensure that clinicians can easily understand the episode window and construction, which is important for the goal of the measure to provide actionable information to clinicians.

The conditions to trigger episodes and the duration of the episode window were established with input from clinician experts in consideration of the goals of the measure to provide actionable information to clinicians about their resource use for a comparable patient cohort. An initial Draft List of Episode Groups and Trigger Codes was posted in December 2016 incorporating input from a Clinical Committee of more than 70 clinicians from over 50 professional societies. Feedback from a four-month public

comment period on that posting was summarized and shared with clinical experts who used the information from the draft list as a starting point and took feedback into consideration along with analyses to help inform discussions (e.g., frequency of services over a period of time extending from the trigger date). This measure was field tested in 2018, as discussed further in Section S.8.3. Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels clinician expert workgroup took field testing feedback into consideration in making refinements to the measure, including feedback on episode exclusions and risk adjustors.

S.8.5. Clinical severity levels *(Detail the method used for assigning severity level and provide rationale for this methodology)*

Clinical severity levels are embedded in the risk adjustment model, as described in Section S.7.5. The model, which is based on the CMS-HCC model and is described in further detail in Section S.8.6, includes variables indicating a patient's health status at the start of the episode. In addition, the risk adjustment model includes stratifications for the sub-groups of the measure for one-level fusion, two-level fusion, and three-level fusion, to ensure that more complex procedures are adjusted for separately. If the procedure occurs inpatient, the risk adjustment model adjusts for the MS-DRG, as there are separate MS-DRGs to indicate spine fusion with complication and comorbidity, with major complication and comorbidity, or without complication and comorbidity/without major complication and comorbidity.

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

S.8.6. Comorbid and interactions *(Detail the treatment of co-morbidities and disease interactions and provide rationale for this methodology.)*

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure accounts for comorbid conditions and interactions by broadly following the CMS-HCC risk-adjustment methodology, which is derived from Medicare Part A and B claims and is used in the Medicare Advantage (MA) program. The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels model includes 79 HCC indicators used in the CMS-HCC V22 2016 model, derived from diagnoses from the patient's Part A and B claims during the 120-day period prior to the episode trigger date, a period that measures conditions that most directly impact patients' health status at the time of the procedure. Episodes where the patient is not enrolled in both Medicare Part A and Medicare Part B for the 120 days prior to the episode are excluded because information on comorbidities for these patients will be incomplete. When applying the CMS-HCC framework to the measure, expected costs are determined by the risk adjustment model separately for each sub-group, which allows the effect of patient health status and demographics on episode spending levels to vary by the sub-groups which reflect the level of the lumbar spine fusion procedure.

Because the relationship between comorbidities' episode cost may be non-linear in some cases (i.e., patients may also have more than one disease during a hospitalization episode), the model also takes into account a limited set of interactions between HCCs and/or enrollment status variables as currently used within the CMS-HCC model. The model includes paired-condition interactions such as chronic obstructive pulmonary disease and congestive heart failure, and interactions between conditions and disability status (e.g., disabled and cystic fibrosis). The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels risk adjustment methodology includes only a limited set of interaction terms for two reasons. First, inclusion of too many interaction terms will over-fit the model. Second, the risk adjustment methodology broadly follows the established CMS-HCC risk-adjustment methodology, which uses similar interaction terms.

The model also includes patient age categories, patient disability status, patient ESRD status, and recent use of long-term institutional care. Additionally, the model includes variables that expert clinician input identified as being important to account for on top of the clinical characteristics already defined via the HCCs, including anticoagulant use, obesity, morbid obesity, smoking, rheumatoid disease, and osteoporosis. The full list of variables used in the risk adjustment model can be found in the Measure Codes List, linked at Section S.1.

Adjustments for Comparability

S.9.1. Inclusion and Exclusion Criteria *Detail initial inclusion/exclusion criteria and data preparation steps (related to clinical exclusions, claim-line or other data quality, data validation, e.g. truncation or removal of low or high dollar claim, exclusion of ESRD patients)*

Included populations:

The cohort for this cost measure consists of patients who are Medicare beneficiaries enrolled in Medicare fee-for-service and who receive a lumbar spine fusion that triggers a Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode. To be included, the patient must have an episode ending within the performance period to ensure that the patient's claims record contains sufficient fee-for-service data both for measuring spending and for risk adjustment purposes.

Excluded populations:

Episodes are excluded for data cleaning and completeness reasons, and they are also excluded to ensure comparability by defining a clinically homogenous group of patients. This can help improve the validity of the cost measure by removing sources of variation outside clinician influence and can prevent unintended consequences of measuring clinician cost performance when treating unique patient populations. The following episodes are excluded, with the rationale for each provided below.

- The patient's death date occurred before the episode ended.

  Episodes where the patient died are excluded as they may not accurately reflect a clinician's performance. These episodes are unusually high-cost, potentially due to costly complications or end-of-life services prior to death, and may not accurately reflect the efficiency of the attributed clinician.

- •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. [12]

  This is a standard exclusion implemented across procedural measures with an inpatient component to identify appropriate procedures. Therefore, episodes with trigger claims outside of the appropriate settings are excluded given the measure's intent to capture only lumbar spine fusion procedures performed in acute inpatient hospitals, outpatient facilities, and/or ambulatory care settings. For this measure, episodes with the retained places of service are also risk adjusted by place of service to reflect cost variation that may exist across different settings.

- Episodes with inpatient procedures without relevant MS-DRG codes.

  Episodes will be excluded if the procedure occurred in the inpatient setting and if its concurrent inpatient stay does not have MS-DRG codes that indicate that the reason for admission was for this procedure (i.e., MS-DRGs 453-455, 459, 460). These cases are excluded to limit the measure to only capture admissions where the reason is for the lumbar spine fusion since patients admitted for other reasons (e.g., lumbar fusion with curvature, malignancy, infections, or extensive fusions which are covered under MS-DRGs 456-458) may have different care needs and distinct costs associated with the admission.

- Episodes where the patient has an osteoporotic compression fractures.

  Episodes where the patient has an osteoporotic compression fracture are excluded because these patients likely require a different set of services for lumbar spine fusion-related care compared to the overall patient population.

- Episodes where the patient has cancer.

  Episodes where the patient has a malignant neoplasm of the vertebral column or a secondary malignant neoplasm of the bone are excluded because the care trajectories are different for these patients (i.e., focused on treating the cancer) than lumbar spine fusions for degenerative spine conditions.

- Episodes where the patient has an infection.

  Episodes where the patient has an infection (e.g. intraspinal abscess and granuloma, osteomyelitis) are excluded because these patients require additional antibiotic treatments and more complex post-operative care (as a result of the infection) that is not comparable to patients who undergo lumbar spine fusions for degenerative conditions.

- Episodes where the patient has scoliosis and/or kyphosis.

  Episodes where the patient has scoliosis and/or kyphosis are excluded because these patients often require different (i.e., longer) fusion techniques and spinal instruments on more segments of the spine than would be required for patients undergoing a lumbar spine fusion for degenerative diseases.

- Episodes where any lumbar fusion with curvature, malignancy, infections, or extensive fusions occurs.

  Any episode where lumbar fusions with curvature, malignancy, infections, or extensive fusion occurs are excluded because these patients represent a different population (e.g., MS-DRGs 456-458) than the population of patients who undergo a lumbar spine fusion for degenerative diseases (i.e., MS-DRGs 453-455, 459 or 460) and are therefore outside of the intended measure scope. Patients in this sub-population require more complex surgeries and post-operative care as well.

- Episodes where the patient has experienced trauma.

  Patients who have experienced trauma are excluded because they have different care needs and trajectories due to their condition. These patients may have severe injuries to nerves and the surrounding structures that require more extensive surgeries than patients who undergo lumbar spine fusions for degenerative diseases.

- Episodes where the patient had a previous spinal fusion, except cervical

  Episodes where the patient had a spinal fusion within the 120 days prior to the episode are excluded (except cervical) because the attributed clinician may have no influence on a previous spinal fusion. These patients may also be more clinically complex with higher costs and rates of complications, so exclusion ensures that clinicians will not be penalized for adverse effects or costly complications from the initial procedure.

- Episodes where the patient is undergoing a redo lumbar fusion.

Episodes where the patient undergoes a redo lumbar spine fusion are excluded to ensure that clinicians would not be held accountable for the adverse effects caused by the previous lumbar spine fusion. Retaining such episodes would put the attributed clinician at risk of being attributed a costly episode (potentially due to treating the complications) where they did not have influence over the procedure/outcomes from the previous lumbar spine fusion.

- Episodes classified as outlier cases.

  To account for limitations of risk adjustment, episodes predicted to have expected costs that are substantially different from observed costs are excluded as outliers. Specifically, episodes with residuals from the risk adjustment model below the 1st percentile and above the 99th percentile are considered outliers and removed from measure calculation.

- The patient has a primary payer other than Medicare for any amount of time overlapping the episode window or in the 120 days prior to the episode trigger day.

  This population is excluded to ensure that we have complete claims data for patients as there may be other claims (e.g., for services provided under Medicare Part C) that we do not observe in Medicare Parts A and B claims data. Including episodes that do not meet this criterion could potentially misrepresent a clinician's resource use. This exclusion also allows us to accurately construct HCCs for each episode by examining the episode's lookback period without missing claims.

- No attributed clinician is found for the episode.

  These episodes are excluded as the measure assesses clinician performance. The measure is intended to assess a homogeneous patient cohort to provide meaningful comparisons between attributed clinicians, so to include these episodes could potentially misrepresent these comparisons.

- The patient's date of birth is missing.

  These episodes are excluded as a data cleaning step.

- The patient's death date occurred before the trigger date.

  These episodes are excluded as a data cleaning step.

- The patient was not enrolled in Medicare Part A and B for the entirety of the 120-day lookback period plus episode window, or is enrolled in Part C for any part of the lookback period plus episode window.

  These episodes are excluded as these patients may receive services not observed in the data. Including these episode could make the attributed clinician appear to have lower cost episodes due to incomplete data.

The rationale and testing results for these exclusions are described further in the testing form (Section 2b2).

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure applies risk adjustment, statistical exclusions, and renormalization to further ensure comparability, described in Step 5 of the construction methodology in Section S.7.2. The risk adjustment approach accounts for patient level variation prior to the episode trigger. Statistical exclusions and renormalizations are engaged during measure construction after excluding outlier episodes to ensure that distributions resulting from outlier exclusions remain true to population averages.

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.

[12] Subsection (d) covers hospitals in the 50 states and D.C. other than: psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer.

S.9.2. Risk Adjustment Type (Select type)

Stratification by risk category/subgroup

If other:

S.9.3. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets)*

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.

S.9.4 Costing method

Detail the costing method including the source of cost information, steps to capture, apply or estimate cost information, and provide rationale for this methodology.

Standardized pricing

The methodology used to payment standardize the Medicare claims used to specify this measure is available for download ("CMS Price (Payment) Standardization") from the following URL: https://www.resdac.org/articles/cms-price-payment-standardization-overview.

S.10. Type of score *(Select the most relevant):*

Ratio

If other:

Attachment:

S.11. Interpretation of Score *(Classifies interpretation of a ratio score(s) according to whether higher or lower resource use amounts is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score, etc.)*

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure score is presented as a dollar figure that represents a clinician's average payment-standardized risk-adjusted cost to Medicare across all Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes attributed to them. A lower measure score indicates that the resource use (observed episode costs) is lower than or similar to expected costs for the care provided for the particular patients and episodes included in the calculation, whereas a higher measure score indicates that the resource use (observed episode costs) is higher than expected for the care provided for the particular patients and episodes included in the calculation.

As a cost measure, this measure on its own does not necessarily by itself reflect quality of care. While it does capture consequences of care by including assigned services during the post-trigger period such as for complications, there are other quality metrics that cannot be captured by a cost measure alone. This measure is most meaningful when presented in part of a program such as MIPS where clinicians are also assessed on quality measures.

S.12. Detail Score Estimation *(Detail steps to estimate measure score.)*

As described in Section S.7.2, the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure score is calculated for each clinician (TIN-NPI) or clinician group (TIN) as follows:

(1) Calculate the ratio of observed to expected episode cost for each episode attributed to the clinician/clinician group.

(2) Calculate the average ratio of observed to expected episode cost across the total number of episodes attributed to the clinician/clinician group.

(3) Multiply the average ratio of observed to expected episode cost by the national average observed episode cost to generate a dollar figure representing risk-adjusted average episode cost.

Reporting Guidelines

This section is optional and will be available for users of the measure as guidance for implementation and reporting.

S.13.1. Describe discriminating results approach

Detail methods for discriminating differences (reporting with descriptive statistics--e.g., distribution, confidence intervals).

The measure is used in the MIPS Cost Performance Category for the CY 2020 performance period onwards. As such, it has not yet been reported as part of MIPS scoring and will be reported later in 2021. While this measure does capture consequences of care such as complications, there are other quality metrics that cannot be captured by a cost measure alone. As such, this measure is most meaningful when reported as part of a program such as MIPS where clinicians are also assessed on quality measures.

While this measure has not yet been reported as a part of MIPS, we expect that the measure, when reported, will provide clinicians with details about their performance including measure score, average costs, and a supplemental granular patient-level file with additional episode details, similar to the MIPS CY 2019 cost measures that were reported in 2020. Additionally, the clinician community has had opportunities to review and become familiar with the measure. During measure development, we conducted national field testing where confidential reports containing cost measure performance on the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure at its draft stage of development (and other episode-based cost measures developed at the same time) were available to clinicians and clinician groups meeting a 10-episode case minimum. The purpose of this field testing was to enable clinicians to become familiar with the measure and to provide feedback on the measure specifications for refinement before CMS considered the measure for use in MIPS. During field testing, a National

Summary Data Report was also posted containing summary statistics on the episode-based cost measures, including information on the distribution of TIN and TIN-NPI level measure scores.

S.13.2. Detail attribution approach

Detail the attribution rules used for attributing resources/costs to providers (e.g., a proportion of total measure cost or frequency of visits during the measure's measurement period) and provide rationale for this methodology.

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode is attributed to clinicians (TIN-NPIs) billing the triggering procedure code. At the clinician group level, an episode is attributed to the TIN if its TIN-NPI(s) are attributed an episode by billing the triggering procedure, and all episodes across the TIN's NPI(s) are aggregated. If the same episode is attributed to more than one NPI within a TIN, this episode is only attributed to the TIN once. MIPS allows for participation at both the TIN and TIN-NPI level, and so this measure can be reported to both individual clinicians and clinician groups. Empirical results on provider performance (e.g., reliability) for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure can be found in the measure testing form (i.e., Section 2a2).

Episodes ending during the performance period are included in a clinician's or clinician group's score. For example, if the performance period is a calendar year, the episode end date (i.e., 90 days after the trigger date) must occur during that calendar year. Requiring episodes to end during the performance period ensures that we have complete claims information for the episode.

S.13.3. Identify and define peer group

Identify the peer group and detail how peer group is identified and provide rationale for this methodology.

Episodes are opened by the presence of trigger codes on Part B physician/supplier claims, so the clinician peer group is limited to those clinicians performing this procedure. This ensures that clinician cost performance for this procedure is being assessed on a homogeneous patient cohort. While this measure was developed for use in MIPS, it can be expanded to other clinician programs.

S.13.4. Sample size

Detail the sample size requirements for reporting measure results.

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure will be reported for TINs and TIN-NPIs with 10 or more episodes. The measure is used in the Merit-based Incentive Payment System (MIPS) for MIPS performance period 2020 onwards.

S.13.5. Define benchmarking and comparative estimates

Detail steps to produce benchmarking and comparative estimates and provide rationale for this methodology.

The measure has not been reported yet, as it is being used in the MIPS cost performance category for the 2020 performance period onwards and will be reported later in 2021.

Reporting this measure as part of the cost performance category helps to measure clinicians' resource use for lumbar spine fusion procedures in the Medicare population, and thereby hold clinicians accountable for their cost effectiveness. There is no reporting/data submission requirement. Combined with measures in the other MIPS performance categories, such as the quality performance category, the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure allows CMS to assess the value of care and incentivize both achievement and improvement in the provision of high-quality, cost-effective care.

Validity – See attached Measure Testing Submission Form

SA.1. Attach measure testing form

2021-04-26-testing-form-l-fusion.docx

[Response Ends]

Measure Number (*if previously endorsed*)**:** N/A
Measure Title: Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure
Date of Submission: **1/8/2021**

Type of Measure:

Measure	**Measure (continued)**
☒ Outcome (*including PRO-PM*)	☐ Composite – *STOP – use composite testing form*
☐ Intermediate Clinical Outcome	☒ Cost/resource
☐ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	*

* Indicates the table cell left intentionally blank

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. ***If there are differences by aspect of testing,*** *(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

[Response Begins]

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for* ***all*** *the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.17*)**	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment	☒ other: Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured;*

e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure uses Medicare Parts A and B claims data maintained by CMS. Part A and B claims data are used to build episodes of care, calculate episode costs, and construct risk adjustors. Data from the EDB is used to determine patient-level exclusions and supplemental risk adjustors, specifically Medicare Parts A, B, and C enrollment, primary payer, disability status, end-stage renal disease (ESRD), patient birth dates, and patient death dates. The risk adjustment model also accounts for expected differences in payment for services provided to patients in long-term care based on the data from the MDS. Specifically, the MDS is used to create the long-term care indicator variable in risk adjustment.

For measure testing, data from the United States Census Bureau American Census, United States Census Bureau American Community Survey (ACS), and Common Medicare Enrollment (CME) are used in the analyses evaluating social risk factors in risk adjustment.

1.3. What are the dates of the data used in testing? Testing includes Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure episodes ending from January 1, 2019, to December 31, 2019. The split-sample reliability analysis also includes episodes ending in the 2018 calendar year. For further details, please see Question 1.7.

1.4. What levels of analysis were tested? (*testing must be provided for* ***all*** *the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item S.20*)**	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other:	☐ other:

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

There were 1,415 clinician group practices (identified by Tax Identification Number [TIN]) and 3,330 practitioners (identified by combination of TIN and National Provider Identifier [NPI]) included in testing of the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure (also referred to as “the Lumbar Spine Fusion Measure”). Clinicians and clinician groups were included if they were attributed 10 or more Lumbar Spine Fusion Measure episodes, as identified in Medicare Parts A and B claims data, ending from January 1, 2019, to December 31, 2019. Episodes were included from all 50 States and D.C. in the following settings: acute inpatient (IP) hospitals, outpatient (OP) facilities, ambulatory/office-based care centers, and ambulatory surgical centers (ASC).

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

There were 49,168 Medicare patients (from 49,671 episodes) included in TIN level testing, and 41,874 patients (from 42,304 episodes) included at TIN-NPI level testing. Lumbar Spine Fusion Measure episodes are triggered by Current Procedural Terminology (CPT) / Healthcare Common Procedure Coding System (HCPCS) codes on Part B Physician/Supplier claims which indicates occurrence of a lumbar spine fusion procedure.

Episodes were included in the sample if they met a set of inclusion criteria (listed below), meant to ensure data completeness and focus the measure on a clinically homogeneous cohort of patients receiving a surgery for lumbar spine fusion. As previously mentioned, a 10 episode case minimum was also applied. These inclusion criteria are listed below:

- The patient had Medicare as their primary payer for the entire episode window, as well as the 120 days prior to the trigger day (the 120-day lookback period).
- The patient was continuously enrolled in Medicare Parts A and B, and not enrolled in Part C, for the entirety of the episode window and the 120-day lookback period.
- The patient date of birth is not missing.
- The patient death date did not occur before the trigger date.
- The patient death date did not occur before episode end.
- The episode can be attributed to at least one main surgeon.
- The episode trigger claim was in an ambulatory/office-based care setting, IP hospital, OP hospital, or ASC based on its place of service.
- If the procedure occurred in an inpatient setting, the inpatient stay occurred in either an acute hospital as defined by subsection (d) or in an acute hospital in Maryland.[1]
- If the procedure occurred in an inpatient setting, the inpatient stay had a relevant MS-DRG code.
- The patient did not have cancer.
- The patient did not have an osteoporotic compression fracture.
- The patient did not have an infection.
- The patient was not undergoing a redo lumbar fusion.
- The patient did not experience trauma due to fracture.
- The patient did not have scoliosis and/or kyphosis.
- The patient did not have a spinal fusion within 120 days prior to the episode, with the exception of cervical spinal fusions.

[1] Subsection (d) covers hospitals in the 50 states and D.C. other than: psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer.

- The episode did not include a procedure with curvature, malignancy, infections, or extensive fusion.
- The episode is not an outlier case.

To ensure that the inclusion criteria listed above do not distort patient characteristics within the measure population, we compared distributions of patient characteristics (age, race, sex, dual eligibility status, income, unemployment, hierarchical condition categories [HCCs]) for patients and episodes before and after applying the inclusion criteria.

Results of this analysis show that the Lumbar Spine Fusion Measure inclusion criteria have a minimal effect on the distribution of patient characteristics within the measure population. Across all demographic categories, the difference in proportion of patients before and after applying inclusion criteria is less than 2.7 percentage points. To illustrate, the measure population is 56.2 percent female before the inclusion criteria is applied, compared with 56.5 percent after criteria is applied at TIN level analysis. This is comparable to TIN-NPI level results where the population is 56.4 percent female after inclusion criteria is applied. When it comes to race categories, the population is 88.5 percent White without inclusion criteria; after inclusion criteria is applied, this statistic is 89.2 percent at TIN level analysis and 89.5 percent at TIN-NPI level analysis. In terms of age, 27.7 percent of the population is between ages 65 and 69 before inclusion criteria is applied, compared with 27.8 percent at TIN level analysis and 27.7 percent at TIN-NPI level analysis after inclusion criteria is applied. Similarly, 27.8 percent of the population is between ages 70 and 74 before inclusion criteria, compared with 29.5 percent at TIN level analysis and 29.7 percent at TIN-NPI level analysis after inclusion criteria. Finally, 18.9 percent of patients are between ages 75 and 79 before inclusion criteria, compared with 19.9 percent at TIN level analysis and 20.1 percent at TIN-NPI level analysis after the criteria is applied.

Full results of this analysis can be seen in Appendix Table 1.6. These results indicate that there is minimal shift in patient characteristics as a results of the inclusion criteria listed in this section.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

The split-sample testing for reliability (described in Section 2a2) includes episodes from calendar years 2018 and 2019. All other testing used the study period of January 1, 2019, to December 31, 2019.

The exclusion analysis (described in Sections 1.6, 2b2, 2b6) used a greater population of episodes without inclusion criteria applied. This includes 3,274 TINs and 11,770 TIN-NPIs, 92,441 patients, and 94,872 episodes. All other testing used the study population as outlined in Section 1.5 and Section 1.6.

1.8 What were the social risk factors that were available and analyzed ? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

The social risk factors analyzed were variables from the ACS, EDB, and CME. All ACS variables were firstly defined at the Census Block Group level and then ZIP code when census block group is missing. Social risk variables analyzed include the following:

1. Income (ACS): Low Income: median income < 33rd percentile nationally; Medium Income: median income in the interval spanning the 33rd percentile to the 66th percentile nationally; High Income: median income > 66th percentile
2. Education (ACS): Education < High School: when % with < high school education is the highest for a given Census Block Group; Education = High School: when % with only high school is the highest; Education > High School: when % with > high school is the highest
3. Employment (ACS): Unemployment Rate > 10%; Unemployment Rate <= 10%
4. Race (EDB): Asian, Black, Hispanic, North American Native, White, and Other
5. Sex (EDB): Female, male
6. Dual status (CME): Full dual, partial dual, non-dual
7. Agency of Healthcare Research and Quality (AHRQ) SES Index: AHRQ index scores are calculated using the AHRQ scoring algorithm and is a continous dependent variable as a replacement of all SES variables. The index includes percentage of households containing one or more person per room, median value of owner-occupied dwelling, percentage of persons below the federal poverty line, median household income, percentage of persons aged ≥ 25 years with at least 4 years of college, percentage of persons aged ≥ 25 years with less than a 12th grade education, and percentage of persons aged 16 or older in the labor force who are unemployed.[2,3]

____

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability Score

Measure reliability is the degree to which repeated measurements of the same entity agree with each other. For measures of clinician performance, the measured entity is the TIN or TIN-NPI, and reliability is

[2] Agency for Healthcare Research & Quality, Centers for Medicare & Medicaid Services, and RTI International. "Creation of New Race-Ethnicity Codes and Socioeconomic Status (SES) Indicators for Medicare Beneficiaries." Research Triangle Park, 2008. https://archive.ahrq.gov/research/findings/final-reports/medicareindicators/index.html

[3] SES Index Score = 50 + (-0.07 * [% of households containing one or more person per room]) + (0.08 * [median value of owner-occupied dwelling, standardized range from 0-100] + (-.010 * [% of persons below the federally defined poverty line]) + (0.11 * [median household income, standardized range from 0-100]) + (0.10 * [% of persons aged ≥ 25 years with at least 4 years of college] + (-0.11 * [% of persons aged ≥ 25 years with less than a 12th grade education]) + (-0.08 * [% of persons aged 16 or older in the labor force who are unemployed])

the extent to which repeated measurements of the TIN or TIN-NPI give similar results. To estimate measure reliability, we used a signal-to-noise analysis.

In line with NQF guidance in the Committee Guidebook, the signal-to-noise analysis seeks to determine the extent to which variation in the measure is due to true, underlying clinician performance, rather than random variation (i.e., statistical noise) within clinicians due to the sample of cases observed. To achieve this, we calculate reliability scores as:

$$R_j = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_{w_j}^2}$$

where $\sigma_{w_j}^2$ is the within-group variance of the mean measure score of clinician *j*, and σ_b^2 is the between-group variance of clinicians in the measure. That is, reliability is calculated as the ratio of between-group variance to the sum of between-group variance and within-group variance. Reliability closer to a value of one indicates that the between-group variance is relatively large compared to the within-group variance, which suggests that the measure is effectively capturing systematic differences between the clinician and their peer cohort.

Split Sample Reliability Testing

This test examines agreement between two measure scores for each TIN or TIN-NPI, calculated from two independent subsets of episodes randomly and evenly split from a larger sample of episodes in 2018 and 2019. For this analysis, two years of data are used to achieve episode volumes per TIN or TIN-NPI that are comparable to episode volumes in a single year, as this measure is calculated and reported for a one-year performance period in MIPS. Good agreement indicates that the measure score is more the result of TIN or TIN-NPI characteristics (i.e., provider care efficiency) rather than statistical noise due to random variation.

Only TIN and TIN-NPIs that met a case minimum of 10 episodes in both samples were included. When creating the split-samples, the larger sample was stratified by calendar year, ensuring that episodes within each calendar year were evenly distributed across the split-samples for each TIN or TIN-NPI. The same methodology was used to calculate performance scores across both split-samples. We then calculated Shrout-Fleiss intraclass correlation coefficients ICC (2,1) between the performance scores to measure reliability. Lower ICC scores indicate less correlation between the two estimates, while higher scores indicate greater agreement, with a score of 1 indicating that the estimates are exactly the same.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Reliability Score

Table 1 presents the distribution of reliability scores for TINs and TIN-NPIs captured in the measure. At a volume threshold of at least 10 episodes, the mean reliability for TINs is 0.78 and for TIN-NPIs is 0.72.

Table 1. Distribution of Reliability Scores for TINs and TIN-NPIs (10 Episode Volume Threshold)

Reporting Level	# of TINs or TIN-NPIs	Mean (Std. Dev.)	Distribution (by Percentile): 10th	Distribution (by Percentile): 25th	Distribution (by Percentile): 50th	Distribution (by Percentile): 75th	Distribution (by Percentile): 90th
TIN	1,415	0.78 (0.11)	0.64	0.69	0.79	0.87	0.92
TIN-NPI	3,330	0.72 (0.08)	0.60	0.65	0.71	0.79	0.84

In response to stakeholder interest in seeing measure reliability for clinician groups of different practice size, Table 2 shows the distribution of reliability scores by the number of TIN-NPIs within a TIN. When examined by number of clinicians within the practice, the average reliability scores increase from 0.71 (1 clinician) to 0.95 (21+ clinicians) for TINs.

Table 2. Distribution of Reliability Scores for TINs by Practice Size (10 Episode Volume Threshold)

# of TIN-NPIs in TIN	# of TINs	Mean (Std. Dev.)	Distribution (by Percentile): 10th	Distribution (by Percentile): 25th	Distribution (by Percentile): 50th	Distribution (by Percentile): 75th	Distribution (by Percentile): 90th
All	1,415	0.78 (0.11)	0.64	0.69	0.79	0.87	0.92
1	220	0.71 (0.09)	0.60	0.64	0.70	0.77	0.83
2-4	595	0.74 (0.09)	0.62	0.67	0.74	0.80	0.87
5-20	562	0.84 (0.09)	0.70	0.79	0.86	0.91	0.94
21+	38	0.95 (0.02)	0.92	0.94	0.96	0.97	0.97

Split-sample Reliability Testing Results

Table 3 presents the Pearson correlation and ICC (2,1) coefficients between the split-sample measures scores. This analysis included 1,176 TINs and 2,328 TIN-NPIs. The ICC coefficient was 0.73 at the TIN-level, and 0.67 at the TIN-NPI level.

Table 3. Split-sample Analysis Results

Reporting Level	# of TINs or TIN-NPIs	Mean Score: Sample 1	Mean Score: Sample 2	Pearson Correlation Coefficient	ICC(2,1) Coefficient
TIN	1,176	1.01	1.01	0.73	0.73
TIN-NPI	2,328	1.00	1.00	0.67	0.67

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i*.e., what do the results mean and what are the norms for the test conducted?*)
Overall, testing results indicated high measure score reliability with an average of 0.78 for TINs and 0.72 for TIN-NPIs at a volume threshold of 10 episodes.[4] Reliability for groups of different practice sizes was high, with mean reliability for the smallest TINs at 0.71.

The split-sample reliability analysis provides further evidence of reliability and repeatability of the performance measure. The ICC(2,1) coefficient was 0.73 for TINs and 0.67 for TIN-NPIs, indicating high or moderate overall reliability for TINs and TIN-NPIs.

The two reliability metrics capture related, but distinct, concepts. Our ICC (2,1) metric will tend to differ from our signal-to-noise metric for two reasons: (i) The denominator of ICC(2,1) includes additional statistical variation arising from true differences in a provider's performance across performance periods; and (ii) The denominator of ICC(2,1) imposes a common variance for the residual across providers, ignoring differences in precision arising from differences in case sizes. Reason (i) makes ICC(2,1) a less relevant metric in this context, since program goals actually require accurately distinguishing systematic performance changes from one period to another, rather than treating them as statistical noise. To avoid this issue, one could alternatively calculate ICC(2,1) using split-half samples from a single performance period. However, this approach also underestimates reliability of the measure for use in the program; in this case, under-estimation occurs because case sizes are artificially cut in half from true case sizes, mechanically reducing precision from the intended application of the measures. We still present both reliability metrics for completeness, but for reasons (i) and (ii), view the signal-to-noise metric as the preferred and more relevant metric.

____

2b1. VALIDITY TESTING

[4] Thresholds for sufficient measure reliability (including the ICC and other reliability methods) vary across sources (see, for example, Portney and Watkins, 2000, for a discussion). Authors provide a range of thresholds; for example, Landis and Koch (1977) classify Kappa statistics in the 0.41-0.60 range as "moderate," 0.61-0.80 range as "substantial," and 0.81-1.00 range as "almost perfect." Koo and Li (2016), on the other hand, classify ICC values in the 0.5-0.75 range as "moderate," 0.75-0.9 range as "good," and above 0.9 as "excellent." Nunnally (1978) is often cited to justify a threshold of 0.7 for "sufficient" reliability. CMS provides the following thresholds: "*We generally consider reliability levels between 0.4 and 0.7 to indicate "moderate" reliability and levels above 0.7 to indicate "high" reliability.*" (Quality Payment Program 2017 Final Rule: 81 FR 77169). The Department of Education provides the following thresholds: "*Reliability of an outcome measure may be established by meeting the following minimum standards: (a) internal consistency (such as Cronbach's alpha) of 0.50 or higher; (b) temporal stability/test-retest reliability of 0.40 or higher; or (c) inter-rater reliability (such as percentage agreement, correlation, or kappa) of 0.50 or higher.*" (What Works Clearinghouse (WWC) Standards Handbook v4, p.78).

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Face validity

The Lumbar Spine Fusion Measure underwent a structured process for gathering detailed input from clinician experts and other stakeholders during measure development. During this process, Acumen incorporated input from (i) the Musculoskeletal Disease Management - Spine Clinical Subcommittee, (ii) the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels workgroup, (iii) a Technical Expert Panel (TEP), (iv) a Person and Family Committee (PFC), and (v) stakeholder feedback from national field testing.

The Clinical Subcommittee comprised 22 members with clinical experience in musculoskeletal disease management of the spine, affiliated with 19 specialty societies. The Clinical Subcommittee provided input at an in-person meeting in April 2018 on the measure scope and composition of a smaller, targeted workgroup to provide detailed input on each aspect of measure specifications.

The Lumbar Spine Fusion Measure workgroup was composed of 13 members, affiliated with 13 specialty societies, including the North American Spine Society, American Association of Neurological Surgeons, and American Medical Association. The workgroup considered empirical analyses and their clinical expertise to provide input during an in-person meeting and several webinars between June and December 2018. Input was gathered in a structured manner including the use of a polling process requiring greater than 60 percent consensus.

The TEP provided high-level guidance and input on the overall direction of measure development and the framework for episode-based cost measures, while the PFC provided input on concepts of healthcare quality and value. In addition, the national field testing feedback period in October and November 2018 offered all stakeholders an opportunity to review and provide input on draft measure specifications and measure feedback reports for attributed clinicians and clinician groups. During this period, 78,221 field test reports for TINs and TIN-NPIs were available for download and review for 11 episode-based cost measures developed throughout 2018.

To gather a formal record of the Lumbar Spine Fusion Measure workgroup's systematic input throughout measure development, workgroup members completed a face validity survey in December 2020 that assessed the measure's ability to fulfill its intent – to meaningfully compare and evaluate clinicians on cost efficiency – based on current specifications. The survey used a Likert scale with values of 1 = Strongly Disagree, 2 = Moderately Disagree, 3 = Somewhat Disagree, 4 = Somewhat Agree, 5 = Moderately Agree, and 6 = Strongly Agree. Overall, 9 of 13 workgroup members completed the survey.

Empirical Validity Testing

We evaluated the empirical validity of the Lumbar Spine Fusion Measure by examining correlation with an NQF endorsed measure of resource use: the Medicare Spending Per Beneficiary (MSPB) Hospital Measure (NQF# 2158), which assesses the risk-adjusted cost to Medicare for services performed by hospitals and other healthcare providers during an MSPB-Hospital episode. Given the focus on resource use across both measures and the substantial role of hospitals in lumbar spine fusion procedures (e.g., in coordination of care during relevant hospital stays), we anticipate that the Lumbar Spine Fusion Measure cost scores for a provider would be consistent with performance on the MSPB Hospital Measure. To assess this consistency, we analyzed the distribution of Lumbar Spine Fusion Measure scores (i.e., observed to expected cost [O/E] ratios) across MSPB performance ratings.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Face Validity

The results of the assessment of face validity indicate that a convened group of experts had high levels of agreement with the measure's ability to provide an accurate reflection of clinician cost. The survey questions and mean rating for each question are provided below:

> **Questions 1-5:** Indicate the extent to which you agree that each of five aspects of the measure specifications - (i) triggers, (ii) exclusion, (iii) service assignment, (iv) episode window, and (v) risk-adjustment variables - helps the measure fulfill its intent to accurately capture a clinician's risk adjusted cost to Medicare for patients who undergo surgery for lumbar spine fusion.

1. **Triggers**

   Responses: 7 members agreed (rating between 4-6), 2 disagreed (rating between 1-3)

   Mean Rating[5]: 4.4 out of 6 (somewhat to moderately agree)
2. **Exclusions**

   Responses: 8 members agreed, 1 disagreed

   Mean Rating: 4.8 out of 6 (somewhat to moderately agree)
3. **Service Assignment**

   Responses: 6 members agreed, 3 disagreed

   Mean Rating: 3.9 out of 6 (somewhat disagree to somewhat agree)
4. **Episode Window**

   Responses: 8 members agreed, 1 members disagreed

   Mean Rating: 4.7 out of 6 (somewhat to moderately agree)
5. **Risk Adjustment Variables**

   Responses: 8 members agreed, 1 disagreed

   Mean Rating: 4.7 out of 6 (somewhat to moderately agree)

[5] The mean rating is a simple average, calculated by multiplying the number of responses for each rating by the rating, and dividing by the total number of responses.

Mean Response Rating: 4.5 out of 6 (somewhat to moderately agree)

The mean rating from these five questions indicates overall consensus agreement on the measure specifications, and reflects the strength of the measure development process, wherein expert clinicians engage with the details of measure design to ensure that each component (e.g., triggers, exclusions, assigned services) facilitates valid clinician performance measurement.

Beyond these five questions, we also asked members to indicate the extent to which they agree that "the scores obtained from the Lumbar Spine Fusion Measure as specified will provide an accurate reflection of the costs for episodes of care, and can be used to distinguish good and poor performance on cost effectiveness." However, we were unable to obtain all the NQF-requested information on this question (e.g., the "degree of consensus" and "any areas of disagreement") ahead of the testing form submission deadline[6]. We intend to follow up with survey respondents to ensure that we receive feedback on this question to supplement existing results on measure validity.

Empirical Validity

Tables 4 and 5 below present Lumbar Spine Fusion Measure cost scores stratified by performance ratings on the MSPB Hospital Measure, grouped to allow for sufficient provider counts in each performance stratification. A performance rating of 0 indicates the lowest score on the MSPB Hospital Measure (i.e., low cost efficiency), while a rating of 10 indicates the highest score (i.e., high cost efficiency). Results are provided at the TIN and TIN-NPI levels, respectively.

Table 4. Lumbar Spine Fusion Measure Scores by MPSB Performance Rating, TIN Level

MSPB Performance Rating	TIN Count	Measure Cost Score (Observed Cost/Expected Cost Ratio): Mean	Measure Cost Score (Observed Cost/Expected Cost Ratio): Std. Dev.	Measure Cost Score (Observed Cost/Expected Cost Ratio): P25	Measure Cost Score (Observed Cost/Expected Cost Ratio): Median	Measure Cost Score (Observed Cost/Expected Cost Ratio): P75
Performance Rating 0	1,372	1.04	0.16	0.95	1.01	1.10
Performance Rating 1	471	1.01	0.13	0.94	0.99	1.06
Performance Rating 2	272	0.99	0.13	0.92	0.97	1.05
Performance Rating 3-4	346	0.98	0.11	0.92	0.97	1.02
Performance Rating 5-10	163	0.96	0.14	0.89	0.94	1.00

[6] NQF Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement, Page 38.

Table 5. Lumbar Spine Fusion Measure Scores by MPSB Performance Rating, TIN-NPI Level

MSPB Performance Rating	TIN-NPI Count	Measure Cost Score (Observed Cost/Expected Cost Ratio): Mean	Measure Cost Score (Observed Cost/Expected Cost Ratio): Std. Dev.	Measure Cost Score (Observed Cost/Expected Cost Ratio): P25	Measure Cost Score (Observed Cost/Expected Cost Ratio): Median	Measure Cost Score (Observed Cost/Expected Cost Ratio): P75
Performance Rating 0	4,565	1.04	0.17	0.93	1.00	1.10
Performance Rating 1	1,830	1.01	0.15	0.92	0.98	1.06
Performance Rating 2	961	0.99	0.15	0.91	0.97	1.04
Performance Rating 3-4	1,297	0.98	0.13	0.91	0.96	1.02
Performance Rating 5-10	517	0.96	0.13	0.89	0.94	0.99

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Face Validity

This measure was assessed by a group of experts. Out of 9 respondents to the survey, substantial majorities (6 to 8 respondents) agreed that each of the measure specifications helps the measure capture clinician cost performance as intended, and that the scores from the measure, as currently specified, provide an accurate reflection of clinician cost effectiveness. This is furthermore reflected in an overall mean response rating of 4.5 out of 6, indicating a fair level of agreement with each of the key measure components. Altogether, these survey results and expert clinician input demonstrate the high face validity of the Lumbar Spine Fusion Measure.

Empirical Validity

As expected, provider cost scores for the Lumbar Spine Fusion Measure decrease (i.e., cost performance improves) with increases in MSPB Hospital Measure performance ratings. For example, in Table 5, as MSPB performance ratings increase (from 0 to 5-10), the O/E ratios in the Mean column decrease from 1.04 to 0.96, indicating a consistent improvement in Lumbar Spine Fusion Measure performance as well. These results provide meaningful external validation of the Lumbar Spine Fusion Measure, and evidence that the measure can accurately distinguish between good and poor cost performance.

____

2b2. EXCLUSIONS ANALYSIS

NA □ no exclusions — *skip to section 2b3*

2b2.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
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.[7]
- 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.

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
Table 6 below presents observed cost statistics and observed to expected cost ratios for the Lumbar Spine Fusion Measure exclusions. Cost statistics are also provided for the episodes included in the measure for comparison, with a 10 episode case minimum at the TIN and TIN-NPI levels. Full results can be seen in Appendix Table 2b2.2.

[7] Subsection (d) covers hospitals in the 50 states and D.C. other than: psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer.

Table 6. Cost Statistics for Lumbar Spine Fusion Measure Exclusions

Exclusion	Observed Cost: Mean	Observed Cost: 10th Percentile	Observed Cost: 90th Percentile	Observed Cost/Expected Cost: Mean	Observed Cost/Expected Cost: 10th Percentile	Observed Cost/Expected Cost: 90th Percentile
Death in Episode	$59,380	$31,322	$94,463	1.06	0.78	1.46
Not in IP, OP, or ASC Setting	NA	NA	NA	NA	NA	NA
Not in Acute Hospital	$42,971	$28,732	$67,801	1.05	0.81	1.52
No Relevant DRG	$12,043	$3,436	$28,855	0.76	0.18	1.00
Cancer	$68,880	$32,104	$105,212	1.06	0.82	1.50
Fracture	$56,745	$29,850	$85,351	1.05	0.82	1.47
Fracture Trauma	$50,958	$29,387	$81,564	1.05	0.79	1.40
Infection	$62,205	$29,569	$104,229	1.07	0.80	1.43
Curvature, Malignancy, Extensive Fusions	$56,818	$31,052	$91,920	1.06	0.82	1.40
Scoliosis or Kyphosis	$56,752	$30,704	$94,364	1.06	0.82	1.39
Prior Spinal Fusion	$49,969	$27,710	$78,638	1.00	0.65	1.42
Redo Lumbar Fusion	$45,120	$28,425	$69,513	1.02	0.80	1.33
Outlier Cases	$67,701	$4,704	$125,651	1.45	0.22	2.73
Included Episodes (TIN level)	$38,663	$28,022	$55,873	0.99	0.81	1.28
Included Episodes (TIN-NPI level)	$38,552	$27,997	$55,471	0.99	0.81	1.27

* indicates 10 or fewer episodes/patients, in line with the CMS cell size suppression policy[8]

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. **Note: If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
The statistical results provide evidence that excluded episodes are not comparable to the overall measure population. Specifically, the distinct cost and risk characteristics of the episode populations in Table 6 substantiate the clinical rationale, outlined in Section S.9.1 of the Intent to Submit, for why these populations should be excluded to prevent unfair comparisons and potential threats to measure validity. Results for each exclusion are discussed further below.

[8] More information on the CMS cell size suppression policy can be found on this ResDAC website: https://www.resdac.org/articles/cms-cell-size-suppression-policy

Episodes ending in death: Results provide evidence that episodes ending in death include care for costly complications or end-of-life services that may not be comparable to the care provided in the final set of episodes included in the measure. To demonstrate, the average observed cost for episodes ending in death is $59,380, more than $20,000 over the average observed cost for included episodes at the TIN and TIN-NPI levels. The average observed to expected cost ratio is also notably higher for these episodes (at 1.06) compared to the included episode population (which is 0.99 at both TIN and TIN-NPI levels).

Episodes where the trigger claim was not in an appropriate setting: Given the measure's intent to capture procedures performed in acute inpatient hospitals, outpatient facilities, and/or ambulatory care settings, episodes with trigger claims outside of these settings are excluded from the measure. Results show that these episodes are distinct from the measure cohort, highlighting the importance of their exclusion to prevent episodes in irrelevant settings from introducing cost variation unrelated to the measure scope and posing a threat to measure validity. Specifically, results show that these episodes have higher mean observed cost and higher mean observed to expected cost ratios. For example, the average observed to expected cost ratio for episodes not in an acute hospital is 1.05, notably higher than the ratio for included episodes (0.99).

Episodes with inpatient stays that do not have a relevant MS-DRG code: Episodes are excluded if the procedure occurred in an inpatient setting, but the inpatient stay did not have a DRG code relevant to lumbar spine fusions. Results provide evidence that these patients are distinct from the overall measure cohort, demonstrating the importance of excluding these episodes so that they do not introduce cost variation unrelated to the measure scope that could pose a threat to measure validity. Specifically, results show that these episodes are substantially less costly than the population of episodes included in the measure. For example, the mean observed cost for episodes without a relevant DRG is $12,043, less than a third of the mean cost for included episodes.

Episodes with curvature, malignancy, infections, or extensive fusions: These episodes are excluded as they represent patient populations that fall outside the measure scope (e.g., patients undergoing a lumbar spine fusion for degenerative diseases specifically). Results show that these episodes tend to be much higher cost than the overall measure cohort, highlighting the importance of their exclusion to ensure these episodes do not introduce cost variation unrelated to the measure scope that could pose a threat to measure validity. To be specific, the average observed cost for these episodes is $56,818, almost $20,000 over the average observed cost for included episodes at the TIN and TIN-NPI levels. The average observed to expected cost ratio is also notably higher for these episodes (at 1.06) compared to the included episode population (which is 0.99 at both TIN and TIN-NPI levels).

Episodes with patients that have the following clinical conditions: cancer, infections, osteoporotic compression fractures, trauma due to fracture, scoliosis, and kyphosis: Analysis results substantiate clinical rationale that these patients require distinct care (e.g., care focused on cancer treatment, different fusion techniques for scoliosis or kyphosis) that would make these episodes not comparable to the final set of included episodes. These episode populations tend to have higher observed costs and higher observed to expected cost ratios compared to included episodes. For instance, the mean observed cost for episodes with cancer patients is $68,880, more than $30,000 over the mean observed cost for included episodes. In another example, the mean observed to expected cost ratio for episodes with infection is 1.07, compared to 0.99 for included episodes.

Episodes with a redo lumbar fusion or where the patient had a prior spinal fusion: These episodes are excluded to ensure that the attributed clinician is not held accountable for outcomes of a previous spinal fusion procedure. Results show that these episodes tend to be higher cost than included episodes, highlighting the importance of their exclusion so as not to introduce cost variation not under the reasonable influence of the attributed clinician that could pose a threat to measure validity. For example, episodes where patients had a prior spinal fusion had a mean observed cost of $49,969, more than $10,000 over the mean cost of included episodes. Episodes with a redo lumbar fusion have a mean cost of $45,120 (compared to $38,663 for included episodes at the TIN level and $38,552 for included episodes at the TIN-NPI level).

Outlier cases: Outliers are excluded from the measure calculation to avoid cases where a few extreme outliers have a disproportionate effect on measure score. These cases have a mean observed cost of $67,701, with a wide range from $4,704 at the 10th percentile to $125,651 at the 90th percentile. The mean observed to expected cost ratio is 1.45, ranging from 0.22 at the 10th to 2.73 at the 99th percentile.

____

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

2b3.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 122 risk factors**

☒ **Stratification by 3 risk categories**

☐ **Other,**

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

Differences in case mix are controlled for using a statistical risk model with 122 risk factors. The risk adjustment model for the Lumbar Spine Fusion Measure broadly follows the CMS-HCC risk adjustment methodology used in the Medicare Advantage (MA) program. Severity of illness is measured using HCCs, indicators of enrollment and long-term care status, and disease interactions. Age is captured in 12 categorical variables. Variables are included for factors that affect resource use identified by expert clinician input as important to account for in risk adjustment for this specific measure population.

The model includes 79 HCC indicators (as specified in the CMS-HCC Version 22 [V22] 2016 model) derived from the patient's Parts A and B claims during the period 120 days prior to the episode trigger. Episodes for patients without a full 120-day lookback period (i.e., without patient enrollment in both Medicare Part A and Medicare Part B for the 120 days prior to the episode trigger) have their episodes excluded from the measure. This 120-day period is used to measure patient health status and ensures that each patient's claims record contains sufficient 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 End-Stage Renal Disease (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 patients who live in the community. These enrollment and long-term care status variables are non-diagnostic measures of severity of illness.

The model also accounts for disease interactions by including interactions between HCCs and/or enrollment status variables similar to the MA model. These interactions are included as the presence of certain comorbidities increases costs in a greater way than predicted by HCC indicators alone.

Beyond the variables outlined above, the Lumbar Spine Fusion Measure risk adjustment model also includes additional factors to further isolate costs that attributed clinicians can reasonably influence, informed by recommendations from the clinician workgroup based on clinical expertise and empirical analysis. These additional risk adjustors capture whether the patient has a history of:

(i) Anterior Interbody Fusion
(ii) Same-Day Anterior and Posterior Lumbar Fusions
(iii) Antiplatelet or Anticoagulant Use
(iv) ASC
(v) Medical Back Problems Hospitalization
(vi) Combined Posterior or Posterolateral and Posterior Interbody Fusion
(vii) Anemia
(viii) Dementia
(ix) Osteoarthritis
(x) Home Hospital Bed
(xi) Home Oxygen
(xii) HOPD
(xiii) Hypertension
(xiv) Inpatient
(xv) Nursing Facility Physician Visits
(xvi) Obesity
(xvii) Osteoporosis
(xviii) Outpatient Office
(xix) Posterior Interbody Fusion
(xx) Rheumatoid Disease
(xxi) Smoking/Nicotine Dependence
(xxii) Walking Aid
(xxiii) Wheelchairs

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 will make O/E abnormally large, do not dominate certain clinicians' final scores. 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 episodes that deviate the most from their expected values in absolute terms. The expected cost after excluding these outliers is again renormalized to similarly ensure that average expected costs are the same after outlier removal.

The risk adjustment model outlined above is performed separately for each of the 3 measure sub-groups which are based on the level of fusion:

(i) One-level Lumbar Fusion

(ii) Two-level Lumbar Fusion

(iii) Three-level Lumbar Fusion

Additional logic and codes used for risk adjustment are in the RA and RA Details tabs of the Measure Codes List File (see S.1.). Appendix Table 2b3.1.1 includes regression coefficients and standard errors for each of the covariates used in the risk adjustment model.

2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

Clinical Factors

In deciding which clinical variables to include in the Lumbar Spine Fusion Measure risk-adjustment model, we considered the following 10 factors that NQF has highlighted as important to consider[9]:

(i) Clinical/conceptual relationship with the outcome of interest

(ii) Empirical association with the outcome of interest

(iii) Variation in prevalence of the factor across the measured entities

(iv) Present at the start of care

(v) Is not an indicator or characteristic of the care provided (e.g., treatments, expertise of staff)

(vi) Resistant to manipulation or gaming

[9] NQF, Committee Guidebook for the NQF Measure Endorsement Process v6.0 (Sept 2019), page 50

(vii) Accurate data that can be reliably and feasibly captured

(viii) Contribution of unique variation in the outcome (i.e., not redundant)

(ix) Potentially, improvement of the risk model (e.g., risk model metrics of discrimination, calibration)

(x) Potentially, face validity and acceptability

To demonstrate, expert clinicians in the Lumbar Spine Fusion Measure workgroup were asked to identify risk factors they understood to have clinical relationships with the cost of care for lumbar spine fusions (related to i and x above). Workgroup members were then provided with empirical analyses evaluating the relative prevalence and impact on cost of these potential risk factors to confirm appropriateness for inclusion in the risk adjustment model (related to ii and iii above). Furthermore, throughout measure development, measure testing is conducted (as discussed in Section 2b3.5) to evaluate the impact of including certain risk factors on the overall risk model (related to ix above). Finally, patient risk factors for each episode are always identified by claim information present at the start or prior to the start of the episode, ensuring that risk is not an indicator of the care provided by the attributed clinician and mitigating risk of manipulation or gaming (related to iv, v, vi, vii above).

Based on these criteria, Lumbar Spine Fusion Measure workgroup members recommended accounting for the following patient variables based on their clinical associations with the Medicare cost of a lumbar spine fusion procedure:

(i) procedure was an anterior interbody fusion, which involves a more invasive surgical approach requiring more complex post-operative care;

(ii) procedure was a posterior interbody fusion, as this requires differing postsurgical care;

(iii) procedure was a posterior or posterolateral fusion as this requires differing post-surgical care;

(iv) procedure was part of a same-day anterior and posterior lumbar fusion, as anterior and posterior fusions performed on the same day involve a more invasive surgical approach requiring more complex post-operative care;

(v) procedure was a combined posterior or posterolateral and posterior interbody fusion, as the combination of both approaches involve more complex surgery and differing post-surgical care;

(vi) place of service as the attributed clinician may not have a choice of setting depending on geography and other factors, and there is a cost differential across settings;

(vii) patient history of anti-platelet medications, which is associated with higher risk of post-surgical bleeding;

(viii) patient history or current use of anticoagulants as these patients will likely require more post-surgical monitoring for the condition(s) that led to anticoagulant therapy;

(ix) patient has hypertension, indicating higher risk of cardiovascular complications from the surgery and higher likelihood of high costs outside of the clinician’s influence;

(x) patient has morbid obesity or obesity, which confers a much higher risk of pulmonary, metabolic, and cardiovascular complications from the surgery and could result in higher costs outside the clinician’s influence;

(xi) patient has osteoporosis, as this indicates higher risk during surgery and may require different approaches and management outside the influence of the attributed clinician;

(xii) patient has rheumatoid disease, as fusions done in the presence of rheumatic disease confer a higher risk of pulmonary and cardiovascular complications from the surgery;

(xiii) patient history of smoking, as smoking confers a higher risk of pulmonary and cardiovascular complications from the surgery;

(xiv) patient has a frailty indicator (i.e., Osteoarthritis, Anemia, Home Oxygen, Walking Aid, Dementia, Skilled Nursing Facility Visit, Wheelchair, Home Hospital Bed) as frailty is an inherent condition of the patient, outside of the influence of the clinician, and confers higher risk of complications during and following surgery, and;

(xv) patient had a recent hospitalization for medical back problems within 120 days of the trigger, as hospitalization for back problems indicates a more severe condition.

Beyond these measure-specific factors, we also included CMS-HCC factors in the measure risk-adjustment model based on previous studies evaluating its appropriateness for use in risk adjusting Medicare claims data. The CMS-HCC model was developed specifically for use in the Medicare population, meaning that it accounts for conditions found in the Medicare population and is calibrated on Medicare Fee-for-Service (FFS) patients. Additionally, the CMS-HCC model is routinely updated for changes in coding practices (e.g., the transition from ICD-9 to ICD-10 codes) and is exhaustive on these code sets, ensuring that clinical risk data are reliably and feasibly captured in the model. Because the CMS-HCC model has already been extensively tested and is used for a large Medicare Part C population, we focus our testing (described in Sections 2b3.5 to 2b3.20) on how the CMS-HCC model was adapted to the Hip Arthroplasty Measure.[10,11,12]

As noted previously (in Section 2b3.1.1), the Lumbar Spine Fusion Measure risk adjustment model is run on episodes stratified into subgroups. Subgroups were selected based on expert recommendation from the measure workgroup, with the goal of ensuring clinical comparability among episodes so that the cost measure fairly compares clinicians with similar patient case-mix. Subgroups furthermore allow the risk-adjustment model to adjust for the subgroup variable itself (i.e., the level of spine fusion) and thus more precisely predict costs for the measure overall. The workgroup recommended the following Lumbar Spine Fusion Measure sub-groups based on the level of the spine fusion:

(i) One-level Lumbar Fusion

(ii) Two-level Lumbar Fusion

(iii) Three-level Lumbar Fusion

This stratification ensures that costs were only compared within a level of spine fusion, and that the measure specifically accounts for this key difference in patient risk when evaluating clinician performance. More information on sub-groups can be found in Section 2b3.9.

Social Risk Factors

[10] In 2018, 20 million patients were enrolled in Medicare Part C plans and incurred $230 billion to cover Medicare Part A and Part B services for Medicare Advantage enrollees (MEDPAC Data Book *Healthcare Spending and the Medicare Program*, June 2019, http://www.medpac.gov/docs/default-source/data-book/jun19_databook_entirereport_sec.pdf?sfvrsn=0)

[11] Pope, Gregory C., John Kautter, Melvin J. Ingber, Sara Freeman, Rishi Sekar, and Cordon Newhart. “Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report.” RTI International: March 2011

[12] “Report to Congress: Risk Adjustment in Medicare Advantage”, *CMS* https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/RTC-Dec2018.pdf.

According to a 2014 National Quality Forum report[13], the mechanisms underlying differences in resource use by socioeconomic status and race are complex and may be impacted by factors such as financial resources, community resources, historical and current discrimination, and reduced access to preventive services. Provider assumptions or implicit biases may impact quality of care for patients of different races. These factors may result in inefficient care, increased disease severity, or greater morbidity, leading to higher Medicare spending for patients depending on socioeconomic status or demographic status. Other social risk factors identified by the literature that can affect resource use include income, insurance (e.g., Medicaid), education, race and ethnicity, sex, social relationships, and residential and community context including rurality. [14,15,16]

Given the conceptual relationship between these social risk factors and resource use, we analyzed the impact of the following patient-level and Census-Block Group-level factors: income, education, employment, race, sex, dual status, and AHRQ Index. These factors are also listed in Section 1.8.

We used the CMS Enrollment Database (EDB), and Common Medicare Environment (CME) to determine dual eligibility, race, and sex. Socioeconomic status was determined by two approaches: a) using income, education and employment status as categorical dependents and b) using Agency of Healthcare Research and Quality (AHRQ) SES Index as a continuous dependent. Both approaches used data from the 2017 American Community Survey (5-year file) by linking episodes to census block groups, and ZIP code when census block group is missing.

Social risk factors were examined relative to the base set of risk adjustment variables from the CMS-HCC V22 2016 model, disability status, ESRD status, interaction variables, and recent long-term care use, and in a step-wise fashion to determine the potential value of each social risk factor considered.

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

☒ **Published literature**

☒ **Internal data analysis**

☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

The literature has extensively tested the use of the HCC model as applied to Medicare claims data. Although the variables in the HCC model were chosen to predict annual cost, CMS has also used this risk adjustment model in a number of other settings (e.g., ACOs, previous physician QRUR programs, and other measures such as the MSPB Hospital Measure [NQF #2158]). Recalling that the risk model relies on the existing CMS-HCC

[13] National Quality Forum. "Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors: Technical Report." National Quality Forum: August 2014

[14] National Academies of Sciences Engineering and Medicine (U.S.). Committee on Accounting for Socioeconomic Status in Medicare Payment Programs, Kwan LY, Stratton K, Steinwachs DM. Accounting for social risk factors in medicare payment: a report of the National Academies of Sciences, Engineering, Medicine. Washington, DC: The National Academies Press; 2017

[15] Assistant Secretary of Health and Human Services for Planning and Evaluation. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. Washington, D.C. December 2020

[16] Medicare Payment Advisory Commission. Beneficiaries Dually Eligible for Medicare and Medicaid. 2018

model, testing results for factors included in the CMS-HCC V22 2016 model can be found in the Pope et al (2011) report and the December 2018 CMS Report to Congress on risk adjustment in Medicare Advantage.[17,18]

Appendix Table 2b3.1.1 includes regression coefficients and standard errors for each of the covariates used in the risk adjustment model.

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

We analyzed race, sex, dual status, income, education, and unemployment as social risk factors (more information on these variables can be found in Section 1.8 of this document). Patient race, sex, and dual status were obtained from the EDB and CME, while information on income, education, and unemployment were obtained from ACS data. Patient episodes without geographic information necessary to obtain ACS data were excluded, approximately 1.2 percent of all episodes[19]. Of the included patient population for the Lumbar Spine Fusion Measure, 57 percent are female and 89 percent have non-dual enrollment status. Full measure population demographics can be found in Appendix Table 2b3.4b.

We examined the impact of including these social risk factors in our risk adjustment model by running goodness of fit tests when different risk factors are added and compared to the base risk adjustment model, where the base risk adjustment model refers to the full set of measure-specific and standard risk adjustment variables from the CMS-HCC V22 2016 model, disability status, ESRD status, interaction variables, and recent long-term care use. To do this, we ran a step-wise regression to include the following additional social risk factors on top of the adapted base CMS-HCC model (Model 1):

- Model 2: sex
- Model 3: dual status
- Model 4: sex + dual status
- Model 5: sex + dual status + race
- Model 6: sex + dual status + income + education + unemployment
- Model 7: sex + dual status + AHRQ SES Index
- Model 8: sex + dual status + race + income + education + unemployment
- Model 9: sex + dual status + race + AHRQ SES Index

[17] Pope, Gregory C., John Kautter, Melvin J. Ingber, Sara Freeman, Rishi Sekar, and Cordon Newhart. "Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report." RTI International: March 2011.

[18] "Report to Congress: Risk Adjustment in Medicare Advantage", *CMS* https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/RTC-Dec2018.pdf.

[19] Due to this exclusion, coefficients and model fit presented for the base model analyzed within the SRF testing will slightly differ to those presented for the model testing conducted in Section 2b3.5.

Results from this stepwise analysis do not support the inclusion of social risk factors into the Lumbar Spine Fusion Measure risk-adjustment model, based on the NQF evaluation criteria as outlined in Section 2b3.3a[20]. Please see below for explanations of analysis results in relation to these criteria:

Empirical association with the outcome of interest: Analysis results indicate that the relationship between social risk factors and measure cost scores is inconsistent across factors and subgroups; and when this relationship is negative, adjusting for SRF may introduce bias into performance measurement. This was determined through analysis of model coefficients and p-values for each of the base and SRF models. Although there were many significant p-values which indicated social risk factors are likely predictive factors of resource use, the directions of the coefficients and relationships with measure scores were inconsistent: while the sign of some social risk factors was positive under certain subgroups (e.g., Asian race and partial dual enrollment variables under the one-level spine fusion subgroup), the sign of these same factors was negative under another subgroup (e.g., Asian race and partial dual enrollment variables under the three-level spine fusion subgroup), suggesting that expected costs would in fact be lower for a patient with high social risk. Incorporating social risk factors into risk adjustment in these latter cases could in fact penalize providers for taking on patients with high social risk, and in turn bias performance measurement under the Lumbar Spine Fusion Measure.

Contribution of unique variation in the outcome (i.e., no redundant): Analysis results also suggested that adding social risk factors to the measure risk-adjustment model had minimal impact on measure performance and was largely redundant with current model prediction. This was determined using two methods: by (i) analyzing differences in percentiles of observed to expected episode cost (O/E) ratios (i.e., "performance percentiles") both with and without social risk factors in the model, and (ii) examining correlations between measure scores calculated with and without social factors. Both of these tests demonstrated a minimal impact on performance – even for providers at high and low extremes of risk - from including social risk factors in the model. Under the first test, this was demonstrated by the fact that the overwhelming majority of providers - 95.0 percent of TINs and 94.5 percent of TIN-NPIs – saw no or minimal change (5 percentiles or less) in performance percentile when social risk factors were added to the model. Furthermore, under the second test, measure scores calculated with and without social factors were highly correlated at both the TIN and TIN-NPI levels, with Spearman correlation coefficients of 0.996 and 0.996, respectively.

Based on these results, we believe the Lumbar Spine Fusion Measure risk-adjustment model appropriately accounts for the impact of social risk factors on cost, and that the model overall appropriately aligns with the aforementioned NQF risk-adjustment evaluation criteria.

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

To analyze the validity of the current risk adjustment model, we examined three analyses: (a) R-squared and adjusted R-squared for the regression models, (b) predictive ratios to examine the fit of the models at different

[20] NQF, Committee Guidebook for the NQF Measure Endorsement Process v6.0 (Sept 2019), page 50

levels of patient complexity, and (c) coefficient estimates, standard errors, and p-values for the risk-adjustment model.

(a) *R-squared and adjusted R-squared* were calculated to analyze the proportion of observed cost variation explained by the risk-adjustment model. Please note that the results of these tests should be evaluated in the broader context of the Lumbar Spine Fusion Measure. First, a valid measure could have a lower R-squared (i.e., the model not explaining much of the observed cost variation) if observed cost (appropriately) varies more with provider performance than patient characteristics, as the model uses patient-level variables. Secondly, this measure utilizes service assignment rules to ensure that only clinically relevant costs are included in the measure. The exclusion of clinically unrelated services may however reduce the explained portion of the cost variance and the model's R-squared, as these services may be well predicted by patient risk factors in the model. In this case too, a low R-squared does not necessarily indicate that a measure is not valid, while a high R-squared does not necessarily indicate the opposite. These results are discussed in Section 2b3.6.

(b) The *predictive ratios* aim to examine the fit of the model at different levels of patient complexity to examine the model's ability to predict both very low and high cost episodes. Specifically, we created a "risk decile" for each episode calculated as the expected cost values from each episode divided by the national average expected cost value. After arranging episodes into deciles based on the risk, we calculated the average predictive ratio for each decile by using the formula of average (expected cost)/average (observed cost) for all episodes in each decile. These are discussed in Section 2b3.8.

(c) *Coefficient estimates, standard errors, and p-values* were produced to consider the extent to which the coefficients for the risk factor covariates are predictive of episode cost. Results for individual risk adjustment variables should be viewed in the context of the entire model, rather than being analyzed individually. For instance, coefficients indicate the incremental effect of a model variable, holding all other variables fixed. As another example, interactions between model variables must be interpreted in concert with the effects of those variables in isolation. Predictive ratios are provided to aid in the overall assessment of the predictive ability of the risk adjustment model. These results are provided in Appendix Tables 2b3.1.1 and 2b3.7.

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

The overall R-squared for the Lumbar Spine Fusion Measure, calculated by dividing explained sum of squares by total sum of squares, is 0.516. The adjusted R-squared is 0.513.

Appendix Table 2b3.1.1 also includes regression coefficients and standard errors for each of the covariates used in the risk adjustment model. More information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.[21]

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

We interpret calibration as how accurately the risk model's predictions match the actual episode cost. We calculate the average O/E cost ratio for each risk decile to demonstrate the model's prediction accuracy for both high and low cost episodes. The average observed to expected cost is generally close to one, 0.99 to 1.01, across risk deciles, indicating that the model is accurately predicting actual episode cost across risk deciles. Full results can be seen in Appendix Table 2b3.7.

[21] Ibid.

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Analysis of predictive ratios by risk decile for the measure shows that the model has consistent ratios across risk score deciles, with each decile having an average ratio between 0.99 and 1.01. Full results can be seen in Appendix Table 2b3.7.

2b3.9. Results of Risk Stratification Analysis:

As mentioned in Sections 2b3.1.1 and 2b3.3a, we stratify Lumbar Spine Fusion episodes into subgroups to ensure that the measure appropriately accounts for differences in patient risk across these populations during clinician performance measurement. Based on this, we would not expect differences in patient complexity (e.g., from different levels of spine fusion) to be reflected in average subgroup measure scores (i.e., O/E cost ratios). In line with this expectation, average O/E cost scores are similar across subgroups (e.g., 1.01 for one- and two-level spine fusions, and 1.00 for three-level spine fusion, at the TIN level), showing the efficacy of these stratifications.

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)
As demonstrated in Sections 2b3.7 and 2b3.8, the average O/E cost ratios for all risk deciles are close to one. These results indicate that the model is accurately predicting spending, regardless of overall risk level. There was no evidence of excessive under- or over-estimation (i.e., expected cost too low or too high relative to observed cost) at the extremes of episode risk.

The R-squared values for the model are in line with values presented in similar analyses of risk adjustment models.[22] As noted in Section 2b3.5, these results should be interpreted alongside the measure context. Specifically, the measure implements detailed service assignment rules designed to only capture clinically related services that would represent meaningful differences in provider cost. As previously mentioned, service exclusions (and other design features based on expert clinician input) improve the validity and actionability of the measure (by removing cost variation outside a clinician's sphere of influence), but tend to reduce its fit statistics (e.g., adjusted R-squared). This is because patient-driven cost variation and excluded services outside the influence of the attributed clinician may be well predicted by patient-level risk factors. Thus, excluding this clinically-unrelated and patient-driven cost variation can reduce the explained portion of the cost variance and the model's adjusted R-squared.

2b3.11. Optional Additional Testing for Risk Adjustment (***not required****, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*) N/A

____

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
We use two methods to identify statistically significant and meaningful differences in the Lumbar Spine Fusion Measure:

[22] Ibid, 6.

- First, we analyzed the distribution of performance scores for the overall measure, as well as for clinicians stratified by meaningful provider characteristics (urban/rural, census division, census region, and the number of episodes attributed to the clinician). The purpose of this analysis is to ensure that there is a sufficiently large difference in measure scores among clinicians to meaningfully determine a difference in performance. In addition, this analysis looks to confirm that the measure behaves as expected with respect to meaningful clinician characteristics.
- In our second test, 95% confidence intervals (CI) were calculated using the variance of the provider mean. We then compared each clinician's 95% CI to the national average measure score to determine if the clinician's performance was significantly different from the national mean. Specifically, clinician performance was deemed to be statistically significantly higher than the national mean if the 95% CI was above the national mean. Conversely, clinician performance was deemed to be significantly lower than the national mean if the 95% CI fell below the national mean. The analysis further confirms that there is a sufficiently large difference in measure scores among clinicians to meaningfully determine differences in performance.

2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The Lumbar Spine Fusion Measure scores have a good deal of variability. For TINs, the standard deviation is 0.09, and 99/1, 90/10, and 75/25 percentile ratios are 1.56, 1.23, and 1.10, respectively. For TIN-NPIs, the standard deviation is 0.10, and 99/1, 90/10, and 75/25 percentile ratios are 1.56, 1.24, and 1.11, respectively. Figure 1 displays the distribution of TINs' mean observed to expected cost ratios across attributed episodes, a direct scalar of provider measure score. Figure 2 displays the same distribution but at a TIN-NPI level.

Figure 1. Distribution of Average Observed to Expected Cost Ratios for TINs

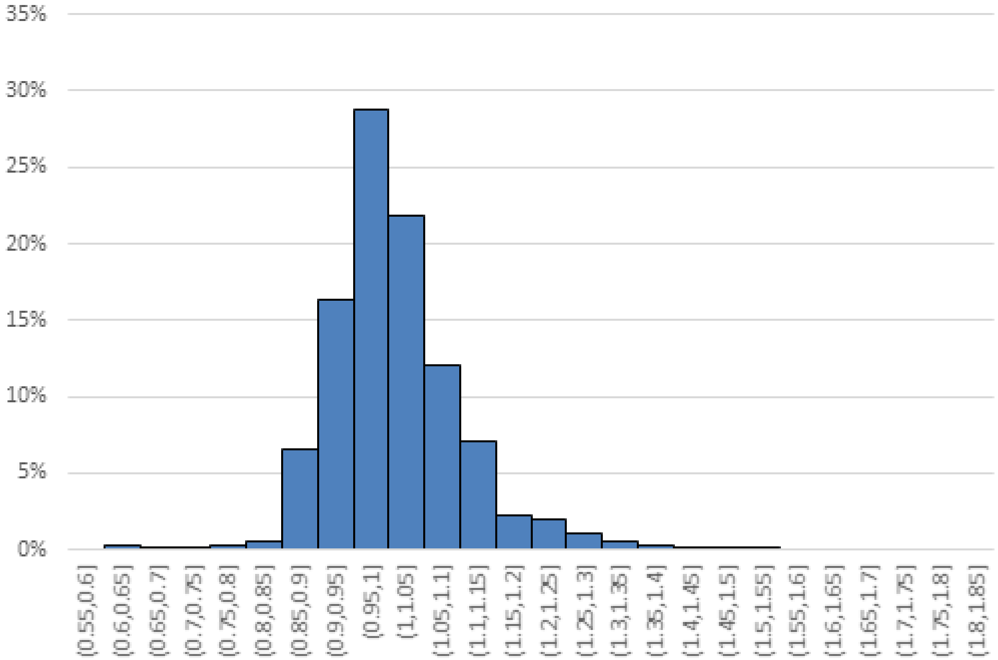
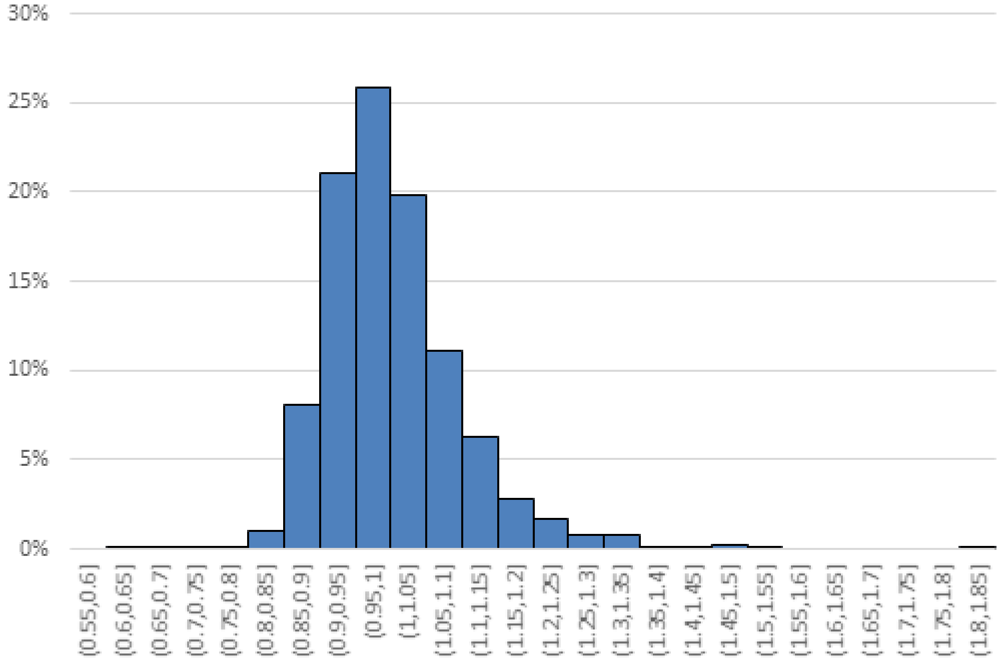


Figure 2. Distribution of Average Observed to Expected Cost Ratios for TIN-NPIs



Analysis results also show that there is not systematic regional differences in clinician score. For instance, at TIN-NPI level of analysis, clinicians in urban areas have a mean performance score of 1.00, which is comparable to the mean score of clinicians operating in rural areas at 0.99. Differences in mean score across four census regions is also limited, ranging from 0.99 in the Midwest region to 1.02 in the Northeast region.

Analysis of clinicians by number of episodes indicates that clinicians with more episodes perform similarly to those who perform fewer procedures. Specifically, at TIN-NPI level of analysis, clinicians with fewest episodes (10-19) had an average score of 1.01, while clinicians with higher case volumes (30-59) had an average score of 0.99. Full results can be seen in Appendix Table 2b4.2.

Due to the high level of reliability of the Lumbar Spine Fusion measure, demonstrated in Section 2a2, small differences in scores can be interpreted as meaningful. This is confirmed by our analysis of statistical significance: 11.45 percent of TINs and 10.63 percent of TIN-NPIs had scores that were statistically significantly higher than the national mean, while 13.22 percent of TINs and 8.68 percent of TIN-NPIs had scores that were statistically significantly lower (Table 7).

Table 7. Proportion of Measure Scores Statistically Significantly Different From the National Average

Provider Level	# of Providers	Statistically significantly lower than national mean: #	Statistically significantly lower than national mean: %	Not statistically significantly different from national mean: #	Not statistically significantly different from national mean: %	Statistically significantly higher than national mean: #	Statistically significantly higher than national mean: %
TIN	1,415	187	13.22%	1,066	75.34	162	11.45

Provider Level	# of Providers	Statistically significantly lower than national mean: #	Statistically significantly lower than national mean: %	Not statistically significantly different from national mean: #	Not statistically significantly different from national mean: %	Statistically significantly higher than national mean: #	Statistically significantly higher than national mean: %
TIN-NPI	3,330	289	8.68%	2,687	80.69	354	10.63

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i*.e., what do the results mean in terms of statistical and meaningful differences?*)
There are clinically and practically significant variation in Lumbar Spine Fusion Measure scores, indicating the measure's ability to capture differences in performance. Our findings regarding variation in measure s cores are consistent with expert clinician input and the face validity rating from expert clinicians that scores obtained from the measure specifications will provide an accurate reflection of the cost of episodes of care, and can be used to distinguish good and poor performance on cost effectiveness (see Section 2b1.2). For example, empirical results indicate that clinicians are not being penalized or rewarded due to risk score decile or type of clinician (e.g., clinicians practicing in rural vs urban settings, or small vs large providers). These suggest that differences in scores are due to meaningful differences in performance, rather than patient or clinician effects. In this way, the measure can capture meaningful differences in resource use and, thus, provide actionable feedback to clinicians on how to improve their performance through care practice changes.

____

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i*.e., what do the results mean and* *what are the norms for the test conducted*)

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Since CMS uses Medicare claims data to calculate the Lumbar Spine Fusion Measure, Acumen expects a high degree of data completeness. To further ensure that we have complete and accurate data for each patient who opens an episode, we exclude episodes where patient date of birth information (an input to the risk adjustment model) cannot be found in the Enrollment Database (EDB), the patient does not appear in the EDB at all, or the patient death date occurs before the episode trigger date. We also exclude episodes where the patient is enrolled in Medicare Part C or has a primary payer other than Medicare in the 120-day lookback period and episode window. In such situations, Medicare Parts A and B claims data may not capture the complete clinical profile for the patient needed in risk adjustment. Furthermore, Parts A and B claims data may not capture all Medicare resource use if some portion of the patient's care is covered under Medicare Part C. These measure exclusions are meant to prevent potential threats to measure validity from missing or incorrect data.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse;* ***if no empirical sensitivity analysis****, identify the approaches for handling missing data that were considered and pros and cons of each*)
The table below presents the frequency of missing data across five relevant measure exclusion criteria:

- Patient date of birth is missing
- Patient death date occurred before the trigger date
- Episode had no main surgeon
- Patient has a primary payer other than Medicare during the episode window or in the 120-day lookback period
- Patient was not continuously enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 120-day lookback period and episode window

Frequency is presented as the number of episodes excluded due to each missing data criteria, as well as the number of TINs and TIN-NPIs who had at least one episode excluded due to the missing data criteria.

Table 8. Missing Data Categories for the Lumbar Spine Fusion Measure

Exclusion	# Episodes	# TINs	# TIN-NPIs
Patient birth date is missing	0	0	0
Patient death before trigger	NA	NA	NA
Episode had no main surgeon	562	396	483
Primary payer other than Medicare	10,571	2,243	6,801

Exclusion	# Episodes	# TINs	# TIN-NPIs
No continuous enrollment in Medicare Parts A and B, or was enrolled in Part C	4,693	1,733	4,560

* indicates 10 or fewer episodes, in line with the CMS cell size suppression policy[23]

2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted;* ***if no empirical analysis****, provide rationale for the selected approach for missing data*)

As the Lumbar Spine Fusion Measure is calculated with Medicare claims data, we expects a high degree of data completeness, which is supported by the limited frequency of missing data for patient birth date and invalid patient death date information above. Additionally, the measure removes patients that may have gaps in the Medicare claims history due to alternate enrollment. These data processing steps ensure that we have complete and accurate information needed to calculate the measure, preventing potential threats to measure validity from missing or incorrect data.

[Response Ends]

Feasibility

F.1. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

F.1.1. Data Elements Generated as Byproduct of Care Processes.

Generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

F.2. Electronic Sources

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

F.2.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*)

ALL data elements are in defined fields in a combination of electronic sources

[23] More information on the CMS cell size suppression policy can be found on this ResDAC website: https://www.resdac.org/articles/cms-cell-size-suppression-policy

F.2.1a. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.

F.2.2. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.

Attachment:

F.3. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

F.3.1. Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation iss ues.

Lessons and associated modifications are categorized into three types: data collection procedures, handling of missing data, and data sampling associated with beneficiaries who died before the measurement period.

[Response Begins]

Data Collection

Acumen receives claims data directly from the Common Working File (CWF) maintained at the CMS Baltimore Data Center. Medicare claims are submitted by healthcare providers to a Medicare Administrative Contractor (MAC), and are subsequently added to the CWF. However, these claims may be denied or disputed by the MAC, leading to changes to historical CWF data. In rare circumstances, finalizing claims may take many months, or even years. As a result, it is not practical to wait until all claims for a given month are finalized before calculating this measure. As such, there is a trade-off between efficiency (accessing the data in a timely manner) and accuracy (waiting until most claims are finalized) when determining the length of the time (i.e., the "claims run-out" period) after which to pull claims data. To determine the appropriate claims run-out period, Acumen has performed testing on the delay between claim service dates and claims data finalization. Based on this analysis, Acumen uses a run-out period of three months after the end of the calendar year to collect data for development and testing purposes. MIPS reporting for this cost measure will be done in line with program reporting.

Missing Data

This measure requires complete beneficiary information, and episodes with missing data are excluded to ensure completeness of data and accurate comparability across episodes. For example, episodes where the beneficiary was not enrolled in Medicare Parts A and B for the 120 days prior to the episode start date are not included in this measure. This enables the risk adjustment model to adjust accurately for the beneficiary's comorbidities using data from the previous 120 days of Medicare claims. Additionally, the risk adjustment model includes a categorical variable for beneficiary age bracket, so episodes for which the beneficiary's date of birth cannot be located are not included in this measure.

Sampling

To further ensure data accuracy and completeness of the sample, beneficiaries who die before the episode start date are not included in this measure. These beneficiaries are excluded to ensure that the sample is representative of the patient population who undergo a lumbar spine fusion procedure.

F.3.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, and algorithm)?

N/A

F.3.3. If there are any fees associated with the use of this measure as specified, attach the fee schedule here. (Save file as: F3_3_FeeSchedule)

[Response Ends]

Usability and Use

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

U.1.1. Current and Planned Use

Specific Plan for Use	Current Use (for current use provide URL)
NA	Payment Program Quality Payment Program Merit-based Incentive Payment System https://qpp.cms.gov/mips/overview

U.1.2. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

[Response Begins]

- Name of program and sponsor: Quality Payment Program (QPP) Merit-based Incentive Payment System (MIPS); Centers for Medicare & Medicaid Services.
- Purpose: The Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) established the Quality Payment Program. Under the QPP, clinicians are incentivized to provide high-quality and high value care through Advanced Alternate Payment Models (Advanced APMs) or MIPS. MIPS eligible clinicians will receive a performance-based payment adjustment to their Medicare payment. This payment adjustment is based on a MIPS final score that assesses evidence-based and practice-specific data across the following categories:
  1. Quality
  2. Improvement activities
  3. Promoting interoperability
  4. Cost

As specified in the CY 2020 Physician Fee Schedule final rule (84 FR 62959 through 62979), this measure will be implemented as part of MIPS beginning in the 2020 MIPS performance year and 2022 MIPS payment year.

- Geographic area and number and percentage of accountable entities and patients included:
  - U.S.
  - The number of clinicians in the QPP varies by performance year. For 2019, there were 954,614 MIPS eligible clinicians receiving a payment adjustment. Of the 954,614 eligible clinicians, 99.99% participated in 2019 with 538,323 clinicians participating in MIPS as individuals or

groups and 416,281 clinicians participating in MIPS through APMs. [9] As clinicians have choices on how to participate in the QPP (e.g., through MIPS or the Advanced APMs, as groups or individuals), the exact number and percentage of clinicians who received a performance score on this measure was confirmed after the end of the performance period.

[9] CMS, "2019 QPP Participation Results Infographic," Quality Payment Program, https://qpp-cm-prod-content.s3.amazonaws.com/uploads/1190/QPP%202019%20Participation%20Results%20Infographic.pdf.

U.1.3. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

N/A

U.1.4. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

N/A

U.2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation. How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Development: Lumbar Spine Fusion Clinician Expert Workgroup

During development, Acumen incorporated expert input from the 13 members of the Lumbar Spine Fusion Clinician Expert Workgroup, who provided detailed feedback on the measure's specifications. Workgroup membership drew from the Musculoskeletal Disease Management – Spine Clinical Subcommittee membership, which had a public call for nominations, as well as additional clinicians identified through stakeholder outreach. Acumen worked with CMS to compose a balanced workgroup reflecting the Musculoskeletal Disease Management – Spine Clinical Subcommittee's input on the types of expertise that would be most relevant to the Lumbar Spine Fusion episode group and on those who would be most likely to be clinicians who would be attributed the measure, such as orthopedic surgeons.

Development: Person and Family Committee

During development, Acumen incorporated person and family engagement (PFE) input from interviews a pool of patients and caregivers called the Person and Family Committee (PFC). PFC members included Medicare beneficiaries and caregiver/family members of a Medicare beneficiary who have lived experience with health care and/or patient advocacy, health care delivery, concepts of value, and outcomes that are important to patients across delivery/disease/episodes of care. PFC members provided feedback on the (i) selection of episode groups for development, and (ii) a broad set of questions around constructing measures that will provide meaningful feedback on clinicians' resource use via service assignment, provider attribution, and episode length.

Development: Field Testing

Acumen and CMS conducted a national field test of 11 episode-based cost measures and two population-level cost measures, including the Lumbar Spine Fusion measure, developed during 2018 for a 35-day comment period (October 3, 2018 to November 5, 2018). We provided Lumbar Spine Fusion Field Test Reports to a sample of eligible clinician groups and clinicians. Each report included information for the Lumbar Spine Fusion measure if the clinician or clinician group was attributed 10 or more episodes. [10] This testing sample was selected to balance coverage and reliability, since a key goal of field testing was to test the measure with as many stakeholders as possible. The number of field test reports shared with the public was:

- Total reports for all measures: 793,842
- Total Lumbar Spine Fusion Field Test Reports: 4,824
- TIN reports: 1,468
- TIN-NPI reports: 3,356

All stakeholders, including those who did not receive a field test report, could review a mock field test report that was posted on the CMS website. Other public documentation posted during field testing included: measure specifications (comprising a Draft Cost Measure Methodology document and a Draft Measure Codes List file), a National Summary Data Report, a Frequently Asked Questions document, and a Fact Sheet. [11] During field testing, Acumen conducted education and outreach activities, including a national webinar, office hours with specialty societies, and Help Desk support. Acumen sought feedback on the reports and measure specifications through an online survey, with the option to attach a comment letter.

Implementation: Pre-Rulemaking and Rulemaking

The Lumbar Spine Fusion measure was implemented in MIPS after going through the pre-rulemaking process and notice-and-comment rulemaking. The measure was submitted to and included in the 2018 Measures Under Consideration (MUC) List. It was then considered by National Quality Forum (NQF)'s Measure Applications Partnership (MAP) Clinician Workgroup and Coordinating Committee in December 2018 and January 2019, respectively.

The measure with was proposed for use in the MIPS cost performance category in the CY 2020 Physician Fee Schedule proposed rule. [12] A National Summary Data Report containing information about the measure performance (e.g., measure score distributions by different provider characteristics) was also publicly posted. [13] The Measure Justification Form that provided results for the testing and evaluation of the Lumbar Spine Fusion measure was also available. Stakeholders submitted comments on the proposed rule during a 60-day public comment period. CMS considered these comments and finalized the measure for use in MIPS from the CY 2020 performance period onwards in the CY 2020 Physician Fee Schedule final rule. [14]

[10] The field test reports were available for download from the CMS Enterprise Portal: https://portal.cms.gov/wps/portal/unauthportal/home/.

[11] The Measure Development Process, Frequently Asked Questions, and Fact Sheet documents are posted on the MACRA Feedback Page: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2018-measure-development-process.pdf; https://www.cms.gov/files/zip/macra-2018-field-testing-materials.zip.

[12] The CY 2020 Physician Fee Schedule proposed rule can be found here: https://www.federalregister.gov/documents/2019/08/14/2019-16041/medicare-program-cy-2020-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other.

[13] CMS, "National Summary Data Report: 11 Episode-Based Cost Measures and Two Revised Cost Measures," (Updated Following Field Testing, June 2019), MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2018-national-summary-data-report.zip.

[14] The CY 2020 Physician Fee Schedule final rule can be found here: https://www.federalregister.gov/documents/2019/11/15/2019-24086/medicare-program-cy-2020-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other.

U.2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

Development: Lumbar Spine Fusion Clinician Expert Workgroup

The workgroup convened for meetings at three points in the process to review empirical analyses, prepared by the Acumen team, and use their clinical expertise to provide detailed input on each component of the measure. Before each meeting, Acumen provided results from empirical analyses and other background materials for members to review. After each meeting, Acumen administered a survey that members completed as a formal way to record consensus on measure specifications that were discussed.

First, the workgroup met at an all-day in-person meeting in June 2018 to discuss measure specifications for all components of the measure. At this meeting, the workgroup provided detailed input on the following: (i) the codes that will be used to open/trigger episodes, (ii) the length of the episode window, (iii) the sub-groups to compare like patients, (iv) the services whose costs are included in the cost measure, (v) the variables to include in the risk adjustment model, and (vi) the measure exclusion criteria. Members reviewed analyses of the utilization and timing of all Medicare Parts A and B services in broad timeframes extending before and after the episode trigger to provide input which services should be included as part of the episode costs. Members also provided clinical input on the particular logic conditions or rules that should be used along with the services, such as requiring additional codes to be present along with the service to ensure clinical relevance, assigning costs for the service if it occurs within a shorter timeframe from the trigger than the overall episode window length, or assigning the service only when accompanied by a particular relevant diagnosis that is newly occurring. Members also reviewed data on frequency and costs associated with sub-populations within the episode group's patient cohort to inform input on risk adjustors and exclusions.

In August 2018, the workgroup convened for a webinar for follow-up discussions on the episode sub-groups, service assignment, risk adjustment, and exclusions. Members provided feedback on feedback on refinements based on testing results for the measure as configured based on input received during the in-person meeting.

After field testing, the workgroup met via webinar November and December 2018 to consider stakeholder feedback received during field testing, refine the measure, and review updated testing results. After meeting, Acumen prepared the final measure specifications documentation reflecting the updates.

Development: Person and Family Committee

The PFC provided input at two points during the measure development. Initial conversations with the PFC focused on the broad concepts of health care quality and value. Subsequent discussions focused on patient and caregiver perspectives on the types of episodes that should be prioritized for development.

In June 2018, the PFC provided input through interviews on pre- and post-trigger services, attribution of clinicians, and services perceived as aiding recovery or helping to avoid unnecessary costs and complications to understand opportunities for improvement. This round of PFC input was broken into several buckets of medical treatments, and input related to scheduled surgeries was relevant for the Lumbar Spine Fusion measure. The input from these discussions was shared with workgroup members for their consideration prior to the workgroup in-person meetings in June 2018.

During the second round of input, PFC members who had specific experience with a lumbar spine fusion participated in in-depth interviews. During these interviews, PFC members considered (i) pre- and post-trigger periods and treatment received therein, (ii) services provided by and costs incurred by various clinicians, including those seen before and after the trigger event, (iii) PFC members' perception of value in health care, and (iv) services perceived as aiding recovery or helping to avoid unnecessary costs and complications. The input from these interviews was shared with the workgroup members who considered these findings, alongside stakeholder feedback from a national field testing period (October 2018) and results of testing analyses, in making refinements to the measures at the webinars in November 2018.

Development: Field Testing

During the feedback period, 20,443 field test reports across all episode-based cost measures, including the Lumbar Spine Fusion measure, were downloaded by 1,542 clinician groups (TINs) and 18,901 clinicians (TIN-NPIs). Stakeholder comments from field testing were summarized for the workgroup to consider in recommending refinements to the measures based on the testing data and feedback.

The following sections offer more details on the contents of each report and describe the education and outreach efforts associated with the field testing feedback period.

Data Provided During Field Testing:

Each Lumbar Spine Fusion Field Test Report contained the following information:

- The clinician or clinician group's measure score with the national average score and percentile rank
- Breakdown of cost measure score by episode sub-group with the national average score
- Episode cost breakdown by Medicare Setting and Service Category to show the average cost per episode and share of services with the certain service (i.e., outpatient evaluation and management, ancillary, hospital inpatient, emergency room, post-acute care, and all other services)
- Breakdown of service utilization and cost by selected clinical categorizations of the service assignment rules associated with episode costs during the window to show the average cost per episode, as well as frequency and cost of different categories of clinical services that are clinically relevant to the episode groups

A mock field test report was posted on the CMS MACRA Feedback webpage during the field testing period. Along with the field test report, attributed clinicians and clinician groups received an episode-level CSV file that included the risk profile of their attributed episodes.

Education and Outreach:

Acumen directly conducted outreach via email to tens of thousands of stakeholders using a stakeholder contact list developed through previous education and outreach and clinician engagement efforts, as well as CMS, QPP, and other available listservs. Examples of the types of emails that were sent include:

- General emails to all our contacts from clinician and healthcare provider organizations. These included contacts we gathered over the course of our measure development work, including contacts directly involved in our work and contacts we compiled from our own research.
- Targeted emails to available contact details linked to a TIN or TIN-NPI that received a field test report.
- Targeted emails to a small number of specialty societies whose members we anticipated would receive a field test report to seek their support in informing their members about field testing.

Acumen and CMS hosted two office hour sessions in October 2018, to provide an overview of field testing to specialty societies, discuss what information their members would be particularly interested in, and answer any questions. Across both office hours sessions, there were 50 attendees.

Acumen and CMS hosted a national field testing webinar on October 9, 2018, to provide an overview of the measures being field tested and the information available for public comment. The webinar consisted of an hour-long presentation, outlining (i) the cost measure development activities, (ii) field testing activities, (iii) how to access and understand the confidential field test reports, and (iv) the contents of the reports. The presentation was followed by a 30-minute Q&A session. [15] There were 381 attendees at this webinar.

An informational post-field testing webinar was held on March 27, 2019, to provide an update on all the measures following field testing. The 60-minute webinar provided an overview of the basics of measure construction, highlighted refinements made after field testing, and provided a summary of testing done on the measures. The presentation was followed by a 30-minute Q&A portion. [16] There were around 400 attendees at this webinar.

Implementation: Pre-Rulemaking

There was a public comment period after the release of the MUC list from December 1, 2018, to December 6, 2018, prior to the MAP Clinician Workgroup Meeting. The MAP Clinician Workgroup met on December 12, 2018, to consider measure specifications and testing updates. Following the release of the Clinician Workgroup's preliminary recommendation, the report was open for a public comment period from December 21, 2018, to January 10, 2019. The MAP Coordinating Committee met on January 22-23, 2019, to consider these comments alongside the MAP Clinician Workgroup's recommendation. Both MAP meetings were open to the public.

Implementation: Rulemaking

During the public comment period for the proposed rule from August 14, 2019, to September 27, 2019, stakeholders could review the proposed rule language, measure specifications, National Summary Data Report, and Measure Justification Forms when submitting comments. CMS conducted email outreach via its listserv to notify stakeholders about the release of the proposed rule.

[15] CMS, "2018 MACRA Cost Measures Field Testing," Quality Payment Program, https://qpp-cm-prod-content.s3.amazonaws.com/uploads/442/2018%20MACRA%20Cost%20Measure%20Field%20Testing%20Webinar_Slides.pdf

[16] CMS, "MACRA Cost Measures Post-Field Testing Webinar," Quality Payment Program, https://qpp-cm-prod-content.s3.amazonaws.com/uploads/521/MACRA%20Cost%20Measures%20Post%20Field%20Testing%20_Slides.pdf.

U.2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1. Describe how feedback was obtained.

The overarching feedback that was received on measure performance and implementation included comments from Workgroup members, PFC members, and the broader stakeholder community. Workgroup members provided feedback via survey after each meeting about the discussed measure specifications, and completed a final face validity survey about their level of agreement with the measure's ability to provide an accurate reflection of costs, and to distinguish good and poor performance. The PFC provided feedback via interviews about health care quality and value and on specific measure components (i.e., defining an episode window, service assignment, and attribution). The broader stakeholder community provided comments during National Field Testing about the measure development components and approach and accessing field test reports and supplemental documentation as well as measure-specific comments. The broader stakeholder community also provided comments during pre-rulemaking and rulemaking processes. The feedback is detailed in sections U2.2.2 and U2.2.3, with references to publicly-available feedback, where appropriate.

Development: Lumbar Spine Fusion Clinician Expert Workgroup

Input from the workgroup was gathered via three post-meeting surveys. 11 out of 13 members completed the first post-meeting survey, 8 out of the 13 members completed the second post-meeting survey, and 9 out of the 13 members completed the third set of post-meeting surveys (from parts one and two of the PFTR webinar).

To gather a formal record of the Lumbar Spine Fusion measure workgroup's systematic input throughout measure development, workgroup members completed a face validity survey in December 2020 that assessed the measure's ability to fulfill its intent – to meaningfully compare and evaluate clinicians on cost efficiency – based on current specifications. Overall, 9 of the 13 workgroup members completed the face validity survey.

Development: Person and Family Committee

In June 2018, 9 PFC members participated in interviews related to medical treatments for scheduled surgeries, including lumbar spine fusions. In September-October 2018, detailed interviews were conducted with 4 PFC members with lived experiences related to a lumbar spine fusion procedure.

Development: Field Testing

In total across measures, Acumen received 67 survey responses and 25 comment letters, including many from specialty societies (e.g., American Medical Association, North American Spine Society, American Association of Neurological Surgeons) representing large numbers of potentially attributed clinicians.

Survey responses and comment letters were collected via an online survey, which contained questions on the measure specifications, as well as questions on the reports themselves and supplemental documentation.

Implementation: Pre-Rulemaking

CMS received 4 comments specifically for the Lumbar Spine Fusion cost measure included in the MUC List released in December 2018. [17] After the MAP Clinician Workgroup meeting in December 2018, there was another public comment period for stakeholders to review their preliminary recommendations. The Lumbar Spine Fusion measure received 2 comments. These public comment periods were facilitated by NQF. Stakeholders were able to submit their comments via the NQF website.

Implementation: Rulemaking

CMS did not receive any specific comments for the Lumbar Spine Fusion cost measure in the CY 2020 Physician Fee Schedule proposed rule; however, CMS received 1 comment that applied to all episode-based cost measures, including lumbar spine fusion. Stakeholders could submit comments through the Federal Register website or via mail.

[17] National Quality Forum, “Measure Applications Partnership Clinician Workgroup Discussion Guide,” (2018), http://public.qualityforum.org/MAP/MAP%20Clinician%20Workgroup/2018-2019%20Clinician%20Workgroup%20Archive/MAP_Clinician_Workgroup_Discussion_Guide.html#COMMENTMUC2018-140MIPS.

U.2.2.2. Summarize the feedback obtained from those being measured.

Development: Lumbar Spine Fusion Clinician Expert Workgroup

During the November 2018 webinar, the workgroup reviewed the measure-specific and cross-cutting field testing feedback, as well as results from empirical analyses. The workgroup recommended revisions to the risk adjustors and exclusions, as described in more detail in U.2.3.

Finally, in the face validity survey, results indicated that there was overall consensus agreement on the measure specifications, and reflected the strength of the measure development process, wherein expert clinicians engaged with the details of measure design to ensure that each component (e.g., triggers, exclusions, assigned services) facilitates valid clinician performance measurement.

Development: Field Testing

The Field Testing Feedback Summary Report presents all feedback gathered during the field testing period. [18] CMS received feedback from 4 commenters for the Lumbar Spine Fusion. Feedback included:

- One stakeholder suggested that some post-trigger services assigned to inpatient medical services and outpatient facility and clinician services should be excluded with the rationale that they are based on diagnoses that remain with the patient in the post-operative period.
- One stakeholder suggested that the post-trigger window may be too long.
- One stakeholder suggested that the cost measure should take into account preventive services, noting that there are guidelines for Acute Low Back Pain recommending a 4-6-week long trial application of manual medicine prior to authorizing a lumbar spine fusion procedure.

- One stakeholder commented that Hierarchical Condition Categories (HCCs) alone may be unable to risk adjust for narrowly defined patient cohorts.

The following list synthesizes the key points that were raised more broadly during the field testing feedback period: Stakeholders provided cross-cutting feedback on risk adjustment variables (e.g., cognitive and functional status, academic medical centers, and socioeconomic status), attribution methodology, episode windows and assigned services, and alignment with cost and quality.

- Stakeholder engagement and involvement remains an important aspect of the measure development process. Stakeholders expressed appreciation for the opportunity to provide feedback during field testing and for CMS' continued efforts to involve them in the measure development process. Commenters also valued the decision to operationalize previously collected feedback, as demonstrated through the addition of measure-specific workgroups to the development process.
- Field test reports present useful information for understanding clinician performance, though reduced complexity could encourage more clinician participation. Stakeholders praised the presentation and content of the field test reports. However, the complexity of the information presented in the reports was a challenge for some stakeholders.
- Improved supplemental field testing materials are helpful but can be further refined. Some stakeholders found the supplemental field testing materials to be informative and thorough, providing useful information on field testing and the specifications of the cost measures. However, many noted that although the materials are comprehensive, they remain lengthy and complex, and they believe the amount of information provided is too overwhelming to be useful.
- Ample time for review of field testing reports and materials is vital to collecting meaningful stakeholder feedback. Some stakeholders suggested the field testing period 4 be extended or kept open, given the large amount and complexity of the information that was presented.
- Transparent Clinical Subcommittee and measure-specific workgroup selection and voting encourages buy-in from stakeholders. Some stakeholders expressed concern with the selection and voting processes for the Clinical Subcommittees and workgroups, highlighting that a transparent approach to member selection would ensure an appropriate mix of specialties and clinician types.
- Field test report access continues to present challenges for stakeholders. Some stakeholders noted that they faced difficulties creating accounts and downloading their field test reports from the CMS Enterprise Portal and these challenges may have negatively impacted the number of clinicians that were able to participate in field testing. Stakeholders urged CMS to communicate directly with clinicians receiving field test reports and to find an alternative for delivering and accessing the reports.

Implementation: Pre-Rulemaking

The MAP gives feedback on performance measures from a wide variety of perspectives, with representatives including "consumers, businesses and purchasers, laborers, health plans, clinicians and providers, communities and states, and suppliers."[19] The Clinician Workgroup specifically aims to ensure, "the alignment of measures and data sources to reduce duplication and burden, identify the characteristics of an ideal measure set to promote common goals across programs, and implement standardized data elements."[20] The MAP voted to conditionally support this measure for rulemaking, conditional on submission to the NQF review and endorsement process.

Implementation: Rulemaking/Public Comment

CMS received comments on the proposed measures during the public comment period for the CY 2020 Physician Fee Schedule proposed rule. There were no measure-specific comments received for Lumbar Spine Fusion. However, CMS received comments about the reliability threshold, the cost category weight, and overall actionability of the episode-based cost measures generally.

For more detailed information on the comments received on the measures as part of the proposed rule public comment period, please see the revised cost measures section in the CY 2020 Physician Fee Schedule final rule for a summary of the public comments received along with CMS responses: https://www.federalregister.gov/documents/2019/11/15/2019-24086/medicare-program-cy-2020-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other.

[18] CMS, "October-November 2018 Field Testing Feedback Summary Report for MACRA Cost Measures," (May 2019), MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2019-ft-feedback-summary-report.pdf.

[19] NQF, "Measure Applications Partnership," https://www.qualityforum.org/Setting_Priorities/Partnership/Measure_Applications_Partnership.aspx.

[20] NQF, "MAP Member Guidebook," (August 2020), http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80515.

U.2.2.3. Summarize the feedback obtained from other users.

Development: Person and Family Committee

During the June 2018 interviews, PFC members provided feedback on the pre-and post-trigger windows and categories of assigned services for scheduled services. For the episode windows, PFC members indicated that (i) an episode should begin when the patient and clinician make the decision to pursue a given treatment plan, (ii) the length of the pre-trigger period should vary based on urgency/severity of the condition and wait times, and (iii) an episode should end when the attributed clinician reports the outcome of the treatment plan to the patient, the patient feels better, and/or the treatment plan ends. For the categories of assigned services, PFC members indicated that (i) services included in the episode should be driven by the treatment plan ordered by the attributed clinician (e.g., imaging, labs) or emergency department personnel, (ii) adherence to the treatment plan aided recovery and prevented complications (e.g., home health care, rehabilitation), and (iii) the use of transitional care services and care coordination improved perceptions of quality care following the procedure (e.g., coordination between home health and primary care providers).

During the September-October 2018 interviews, PFC members provided input on the pre-and post-trigger window services, their care team, and the value and quality of their care. PFC members reported having undergone previous, less invasive procedures to address disc degeneration. In the pre-trigger period, PFC members reported receiving imaging services to monitor disc degeneration or assess various treatment options. PFC members also reported receiving services primary care provider and were then referred to an orthopedist for additional services. Many PFC members reported receiving anesthesia-related services prior to the surgery. Following the surgery, some PFC members worked with the surgeon or nurse in planning for physical therapy or rehabilitation. PFC members reported receiving physical therapy services in the post-trigger period, with some receiving services in an inpatient rehabilitation facility, and others in outpatient facilities or at home. Some PFC members reported that the quality of care could have been improved following the surgery to prevent adverse effects, such as post-operative mobility and nerve damage.

Implementation: Pre-Rulemaking

The MAP recognized the importance of cost measures to the MIPS program and conditionally supported this measure pending NQF endorsement. MAP noted that CMS and the Cost and Efficiency Standing Committee should continue to evaluate the risk adjustment model of this measure and consider whether there is need to account for social risk factors in the model. MAP also noted that review of the measure should ensure an appropriate attribution methodology and that the measure adequately considers the issue of small numbers. MAP noted ensuring that cost measures truly address factors within a clinician´s reasonable influence. MAP noted that cost measures should continue surveillance for unintended consequences such as stinting of care and reduced quality of care. MAP noted that cost measures should be paired with balancing measures (e.g., quality, efficiency, access, and appropriate use measures) as one way to safeguard against these issues. MAP recognized a need for continuous feedback and testing of measures as they are implemented. Finally, MAP noted a need to provide greater education on these measures as well as for greater transparency of the measure specifications and testing results. [21]

[21] NQF, MAP Clinicians 2019 Considerations for Implementing Measures Final Report," (March 2019), https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=89597

U.2.3. Describe how the feedback described in 4a2.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not

Development: Person and Family Committee

Input gathered from both rounds of interviews was shared with workgroup members. Specifically, this input informed the workgroup's discussions about the categories of services to assign in the pre-and post-trigger windows and provider attribution or who was involved in the care team, which are described in more detail below.

The workgroup recommended assigning pre-operative testing and services that PFC members mentioned, such as testing, imaging, and related anesthesia and pain management. The workgroup also recommended assigning post-operative services, like post-acute care and rehabilitation, which most PFC members reported receiving after the procedure. PFC members indicated that their care team mainly included the orthopedic surgeon and team at the hospital and the primary care provider, and sometimes a nurse who provided patient education; PFC member input is supported by testing results showing that orthopedic surgeons are the frequently attributed specialty.

Development: Field Testing

After completing field testing, Acumen compiled the feedback provided through the survey and comment letters into a measure-specific report, which was then provided to the Lumbar Spine Fusion Clinician Expert Workgroup, along with empirical analyses to inform their discussion and evaluation of any refinements needed to ensure that the measure is capturing what it was intended to capture.

In addition to the measure-specific comments received during field testing, the workgroup also considered empirical analyses, discussed pending items from previous webinars, and considered cross-measure field testing feedback, and voted to recommend the following refinements, which were implemented [22]:

- Exclude any episode with a Spinal Fusion Except Cervical within 120 days prior to the episode
- Exclude any lumbar spine fusions procedures that have diagnosis codes within MS-DRGs 456-458 (Spinal Fusion with curvature, malignancy, infections, or extensive fusions)
- Add measure-specific risk adjustors for the following frailty variables:
    - Osteoarthritis
    - Anemia
    - Home Oxygen
    - Walking Aid
    - Dementia
    - Skilled Nursing Facility Visit
    - Wheelchair
    - Home Hospital Bed

- ○ Add a measure-specific risk adjustor for a recent hospitalization under MS-DRG 551: Medical Back Problems within 120 days before the trigger

Implementation: Rulemaking/Public Comment

During the public comment period for the CY 2020 Physician Fee Schedule proposed rule, stakeholders submitted comments on the proposed episode-based cost measures. After receiving public comments, Acumen reviewed and evaluated the proposed updates. While we received feedback on the proposed measures generally, as described in Section U.2.2.2, there was no measure-specific feedback received on the specifications of this measure. Therefore, the measure was finalized as proposed.

[22] CMS, "October-November 2018 Field Testing Feedback Summary Report for MACRA Cost Measures," (May 2019), MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2019-ft-feedback-summary-report.pdf.

U.3.1. Progress on Improvement. (Not required for initial endorsement unless available.) Performance results on this measure (current and over time) should be provided in IM.1.2 and IM.1.4.

Discuss:

- **Purpose Progress (trends in performance results)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

U.3.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

N/A

U.4.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

N/A. There were no unexpected findings during the development and testing of this measure.

U.4.2. Please explain any unexpected benefits from implementation of this measure.

N/A. There were no unexpected findings during the development and testing of this measure.

[Response Ends]

Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

H.1. Relation to Other NQF-endorsed Measures

If there are related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

H.1.1. List of related or competing measures (selected from NQF-endorsed measures)

H.1.2. If related or competing measures are not NQF endorsed please indicate measure title and steward.

N/A. There are no related or competing measures that are non-NQF-endorsed cost measures with the same focus and/or the same target population submitted to NQF or implemented in MIPS.

H.2. Harmonization

[Response Begins]

H.2.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications completely harmonized?

No

H.2.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

N/A

H.3. Competing Measure(s)

H.3.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

N/A. There are no competing NQF-endorsed or non-NQF-endorsed cost measures that address the same measure focus and target population.

[Response Ends]

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services

Co.2 Point of Contact: Ronique, Evans, Ronique.Evans1@cms.hhs.gov, 410-786-3966-

Co.3 Measure Developer if different from Measure Steward: Acumen, LLC

Co.4 Point of Contact: N/A, N/A, macra-cost-measures-info@acumenllc.com, 650-558-8882-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

Lumbar Spine Fusion Clinician Expert Workgroup Members:

Anand Rughani, Congress of Neurological Surgeons

Byron Schneider, Spine Intervention Society

David Seidenwurm, American College of Radiology

Erica Bisson, North American Spine Society

Gregory Nicola, American College of Radiology

Heather Smith, American Physical Therapy Association

Jay Nathan, American Association of Neurological Surgeons

Jonathan Gal, American Society of Anesthesiologists

Kimberly Lenington, American Occupational Therapy Association

Mohamad Bydon, Congress of Neurological Surgeons

Morgan Lorio, International Society for the Advancement of Spine Surgery

Peter Sanderson, American Medical Association

Philip Schneider, North American Spine Society

The Lumbar Spine Fusion Clinician Expert Workgroup is composed from the larger Musculoskeletal Disease Management – Spine Clinical Subcommittee. The composition list of the Clinical Subcommittee is included in the Episode-Based Cost Measures Development Process document. [23]

[23]CMS, "Measure Development Process," (October 2018), MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2018-measure-development-process.pdf

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:**

Ad.3 **Month and Year of most recent revision:**

Ad.4 **What is your frequency for review/update of this measure?**

Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**