

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

Measure Justification Form

June 2019



Contents

1.0	Introduction	4
1.1	Project Title and Overview	4
1.2	Measure Name	4
1.3	Type of Measure	4
2.0	Importance.....	5
2.1	Evidence to Support the Measure Focus	5
2.1.1	Measure Description	5
2.1.2	Evidence for Measure Focus	5
2.2	Performance Gap	6
2.2.1	Rationale	6
2.2.2	Performance Scores.....	7
3.0	Scientific Acceptability	8
3.1	Data Sample Description	8
3.1.1	Type of Data Used for Testing.....	8
3.1.2	Specific Dataset Used for Testing	8
3.1.3	Dates of the Data Used in Testing.....	8
3.1.4	Levels of Analysis Tested.....	8
3.1.5	Entities Included in the Testing and Analysis.....	8
3.1.6	Patient Cohort Included in the Testing and Analysis.....	8
3.1.7	Sample Differences	9
3.1.8	Social Risk Factors Included in Analysis	10
3.2	Reliability Testing.....	10
3.2.1	Level of Reliability Testing	10
3.2.2	Method of Reliability Testing	10
3.2.3	Statistical Results from Reliability Testing	11
3.2.4	Interpretation	12
3.3	Validity Testing	12
3.3.1	Level of Validity Testing.....	12
3.3.2	Method of Validity Testing	12
3.3.3	Statistical Results from Validity Testing	14
3.3.4	Interpretation	15
3.4	Exclusions Analysis	15
3.4.1	Method of Testing Exclusions.....	15
3.4.2	Statistical Results from Testing Exclusions.....	17
3.4.3	Interpretation	18
3.5	Risk Adjustment or Stratification	19
3.5.1	Method of Controlling for Differences	19
3.5.2	Conceptual, Clinical, and Statistical Methods	21
3.5.3	Conceptual Model of Impact of Social Risks.....	22
3.5.4	Statistical Results	22
3.5.5	Analyses and Interpretation in Selection of Social Risk Factors.....	22
3.5.6	Method for Statistical Model or Stratification Development.....	23
3.5.7	Statistical Risk Model Discrimination Statistics	24
3.5.8	Statistical Risk Model Calibration Statistics	24
3.5.9	Statistical Risk Model Calibration – Risk Decile.....	24
3.5.10	Results of Risk Stratification Analysis.....	24
3.5.11	Interpretation.....	25
3.6	Identification of Meaningful Differences in Performance.....	25

3.6.1 Method	25
3.6.2 Statistical Results	26
3.6.3 Interpretation	26
3.7 Missing Data Analysis and Minimizing Bias.....	26
3.7.1 Method	26
3.7.2 Missing Data Analysis	27
3.7.3 Interpretation	27
4.0 Feasibility.....	28
4.1 Data Elements Generated as Byproduct of Care Processes	28
4.2 Electronic Sources	28
4.3 Data Collection Strategy	28
4.3.1 Data Collection Strategy Difficulties.....	28
5.0 Usability and Use	29
5.1 Use	29
5.1.1 Current and Planned Use	29
5.1.2 Feedback on the Measure and Development Process	29
5.2 Usability	33
5.2.1 Improvement	33
5.2.2 Unexpected Findings.....	33
5.2.3 Unexpected Benefits	33
6.0 Related and Competing Measures	34
6.1 Relation to Other Cost Measures	34
6.2 Harmonization.....	34
6.3 Competing Measures	34
Contact Information	35

1.0 Introduction

This Measure Justification Form (MJF) provides results for the testing and evaluation of the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure. The MJF is intended to provide detailed information about the testing conducted on this measure, and accompanies the Measure Methodology and Measure Codes List file, which together, comprise the specifications for this cost measure.¹

1.1 Project Title and Overview

The Centers for Medicare & Medicaid Services (CMS) has contracted with Acumen, LLC to develop care episode and patient condition groups for use in cost measures to meet the requirements of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). The contract name is “MACRA Episode Groups and Cost Measures.” The contract number is HHSM-500-2013-13002I, Task Order HHSM-500-T0002.

1.2 Measure Name

Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Episode-Based Cost Measure

1.3 Type of Measure

Cost/Resource Use

¹ CMS, Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure Methodology,” *MACRA Feedback Page*, <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2019-revised-ebcm-measure-specs.zip>. CMS, “Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure Codes List,” *MACRA Feedback Page*, <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2019-revised-ebcm-measure-specs.zip>.

2.0 Importance

2.1 Evidence to Support the Measure Focus

2.1.1 Measure Description

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode-based cost measure (also referred to as “the Lumbar Fusion measure”) evaluates clinicians’ risk-adjusted cost to Medicare for beneficiaries who receive a lumbar fusion. The cost measure score is a clinician’s average risk-adjusted cost for the episode group 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 the 30 days prior to the clinical event that opens or ‘triggers’ the episode, through 90 days after the trigger. Beneficiary populations eligible for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure include Medicare beneficiaries enrolled in Medicare Parts A and B during the performance period.

2.1.2 Evidence for Measure Focus

Policymakers contend that an estimated 80 percent of overall health care costs are attributable to decisions made by clinicians.² However, these same clinicians are often unaware of how their care decisions influence the overall costs of care. One of the goals for using cost measures is to help inform clinicians on the costs attributable to their decision-making, as well as the total cost of their patient’s care. A cost measure offers opportunity for improvement if clinicians can exercise influence on a significant share of costs during the episode, or if lower spending and better care quality can be achieved through changes in clinical practice.

According to the literature and previous feedback received through stakeholder input activities, this measure represents an area where there is opportunity for improvement. An opportunity for improvement for lumbar fusion for degenerative disease exists within a primary performance gap: mitigation of complications, especially wound complications, which increases the risk for readmission.

Medicare beneficiaries have been undergoing elective spine surgery for degenerative changes at increasing rates, and with this has come increasing rates of complications and costs associated with these complications.^{3, 4, 5} Compared to other lumbar spine surgeries such as laminectomies or discectomies, lumbar fusion is associated with greater complication rates due to a variety of factors, including its greater complexity, more extensive dissection, prolonged operative periods, greater risk of intraoperative blood loss, and implant/instrumentation failure, requiring greater health care resource use.^{6, 7} Lumbar fusion surgery can be categorized by

² Fred, Herbert L. “Cutting the Cost of Health Care: The Physician’s Role.” Texas Heart Institute Journal, vol. 43, no. 1, 2016, pp. 4 – 6.

³ Puvanesarajah, V., B. C. Werner, et al. “Morbid Obesity and Lumbar Fusion in Patients Older Than 65 Years: Complications, Readmissions, Costs, and Length of Stay.” [In eng]. Spine (Phila Pa 1976) 42, no. 2 (Jan 15 2017): 122-27.

⁴ Buser, Z., B. Ortega, A. 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.

⁵ Rajaei, S. S., H. W. Bae, et al. “Spinal Fusion in the United States: Analysis of Trends from 1998 to 2008.” [In eng]. Spine (Phila Pa 1976) 37, no. 1 (Jan 1 2012): 67-76.

⁶ Kalakoti, P., S. Missios, et al. “Inpatient Outcomes and Postoperative Complications after Primary Versus Revision Lumbar Spinal Fusion Surgeries for Degenerative Lumbar Disc Disease: A National (Nationwide) Inpatient Sample Analysis, 2002-2011.” [In eng]. World Neurosurg 85 (Jan 2016): 114-24.

⁷ Deyo, R. A., S. K. Mirza, et al. “Trends, Major Medical Complications, and Charges Associated with Surgery for Lumbar Spinal Stenosis in Older Adults.” [In eng]. JAMA 303, no. 13 (Apr 7 2010): 1259-65.

invasiveness, and studies have shown that risk for life-threatening complications was higher with increasing surgical invasiveness.⁸ One study found that the risk-adjusted estimated incremental cost of each complication among Medicare beneficiaries exceeded \$10,000.⁹

Occurrence of complications also contribute to increased risk of readmission. A 2017 study of patients in New York State who underwent lumbar fusion found 25 percent were readmitted within 90 days, with the average time to readmission being 7 days. The most common complications were wound complications at 3.7 percent and wound infections at 3.1 percent.¹⁰ Other studies have similarly found wound complications among the most common complications following lumbar fusion.^{11, 12} Given the impact of surgical complications on resource use, mitigation of these complications provide an area of opportunity for improvement, with potential improvement in care quality and cost savings. One study found that an opportunity for reducing complications exist with intense presurgical planning, medical optimization, utilization of minimally invasive approaches, and adequate communications with general practitioners. By implementing these strategies, there is potential for reducing readmissions as a result of complications.¹³

2.2 Performance Gap

2.2.1 Rationale

Between 2006 and 2012, over 6 million Medicare patients were diagnosed with lumbar degenerative conditions,¹⁴ and lumbar spine procedures are increasingly used in elderly 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 percent in the incidence of fusion procedures within one year of diagnosis.¹⁵ Based on a review of the Medicare Provider Analysis and Review file, total spending on lumbar fusion surgery is also one of the highest admission expenditures in the Medicare program, costing over \$3.6 billion dollars in 2013.¹⁶ The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode-based cost measure was recommended for development by an expert clinician committee—the Musculoskeletal Disease Management - Spine Clinical Subcommittee—because of its high impact in terms of patient population and Medicare spending, and the

⁸ Ibid.

⁹ Culler, S. D., D. S. Jevsevar, 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.

¹⁰ Baaj, A. A., G. Lang, et al. "90-Day Readmission after Lumbar Spinal Fusion Surgery in New York State between 2005 and 2014: A 10-Year Analysis of a Statewide Cohort." [In eng]. *Spine (Phila Pa 1976)* 42, no. 22 (Nov 15 2017): 1706-16.

¹¹ Deyo, R. A., S. K. Mirza, et al. "Trends, Major Medical Complications, and Charges Associated with Surgery for Lumbar Spinal Stenosis in Older Adults." [In eng]. *JAMA* 303, no. 13 (Apr 7 2010): 1259-65.

¹² Martin, B. I., S. K. Mirza, et al. "Hospital and Surgeon Variation in Complications and Repeat Surgery Following Incident Lumbar Fusion for Common Degenerative Diagnoses." [In eng]. *Health Serv Res* 48, no. 1 (Feb 2013): 1-25.

¹³ Baaj, A. A., G. Lang, et al. "90-Day Readmission after Lumbar Spinal Fusion Surgery in New York State between 2005 and 2014: A 10-Year Analysis of a Statewide Cohort." [In eng]. *Spine (Phila Pa 1976)* 42, no. 22 (Nov 15 2017): 1706-16.

¹⁴ Buser, Z., B. Ortega, 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.

¹⁵ Ibid.

¹⁶ Culler, S. D., D. S. Jevsevar, 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.

opportunity for incentivizing cost-effective, high-quality clinical care in this area. Based on the initial recommendations from the Clinical Subcommittee, the subsequent measure-specific workgroup provided extensive, detailed input on this measure.

2.2.2 Performance Scores

Performance scores are provided for 1,440 clinician group practices (identified by Tax Identification Number [TIN]) and 3,286 practitioners (identified by combination of TIN and National Provider Identifier [NPI]). These counts represent attributed clinicians and clinician groups billing Part B Physician/Supplier claims under a Merit-based Incentive Payment System (MIPS) eligible clinician specialty, and do not reflect other MIPS eligibility criteria (e.g., Advanced Alternative Payment Model participation). This table uses a testing volume threshold of 10 episodes.

Table 1: Distribution of Performance Scores

Metric	TIN	TIN-NPI
Mean score	\$36,631	\$36,537
Standard deviation	\$3,513	\$3,669
Score IQR	\$3,948	\$4,161
Score percentile		
10 th	\$32,878	\$32,630
20 th	\$34,011	\$33,695
30 th	\$34,803	\$34,546
40 th	\$35,530	\$35,335
50 th	\$36,282	\$36,062
60 th	\$37,018	\$36,871
70 th	\$37,835	\$37,779
80 th	\$39,069	\$39,028
90 th	\$41,033	\$41,316

3.0 Scientific Acceptability

3.1 Data Sample Description

3.1.1 Type of Data Used for Testing

Medicare administrative claims, Long-Term Minimum data set (MDS), enrollment database (EDB), and Common Medicare Environment (CME)

3.1.2 Specific Dataset Used for Testing

The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure uses Medicare Part A and Part 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 adjusters. Data from the EDB are used to determine beneficiary-level exclusions and supplemental risk adjusters, specifically Medicare Parts A, B, and C enrollment, primary payer, disability status, end-stage renal disease (ESRD), beneficiary birth dates, and beneficiary death dates. The risk adjustment model also accounts for expected differences in payment for services provided to beneficiaries 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 American Census, American Community Survey (ACS), and CME are used in analyses evaluating social risk factors in risk adjustment.

3.1.3 Dates of the Data Used in Testing

The measurement period includes Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes ending from January 1, 2017 through December 31, 2017.

3.1.4 Levels of Analysis Tested

Individual clinician (identified by combination of TIN and NPI) and clinician group/practice (identified by TIN).

3.1.5 Entities Included in the Testing and Analysis

1,440 clinician group practices and 3,286 practitioners were included in the analyses. Clinicians and clinician groups were included in testing if they were attributed 10 or more Lumbar Fusion episodes during the measurement period. Episodes from all 50 States and D.C. in the following settings were included: ambulatory surgical centers (ASC), hospital outpatient departments (HOPD), and acute inpatient (IP) hospitals.

3.1.6 Patient Cohort Included in the Testing and Analysis

48,413 Medicare beneficiaries (from 48,870 episodes) were included in TIN level testing and analysis, and 41,219 beneficiaries (from 41,622 episodes) were included in TIN-NPI level measure testing.

The beneficiary population eligible for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure calculation consists of Medicare beneficiaries enrolled in Medicare Parts A and B (but not Part C) who received lumbar fusion during the measurement period as identified by episode trigger Current Procedural Terminology/Healthcare Common Procedure Coding System (CPT/HCPCS) codes on Part B Physician/Supplier claims. Beneficiaries and their episodes were included in the sample if they met a set of inclusion criteria (listed below) meant to ensure completeness of data and to focus the measure on a clinically homogeneous cohort of patients receiving lumbar fusions.

The inclusion criteria are:

- The beneficiary has 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 beneficiary 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 beneficiary has a sufficient 120-day lookback period.
- The beneficiary date of birth is not missing.
- The beneficiary death date did not occur before episode end.
- The beneficiary does not have a diagnosis of certain cancers related to bone on the trigger claim.
- The beneficiary does not have an osteoporotic compression fracture.
- The beneficiary does not have a lumbar fusion with curvature, malignancy, infections, or extensive fusions.
- The beneficiary does not have an infection diagnosis on the trigger claim.
- The beneficiary is not undergoing a redo lumbar fusion.
- The beneficiary does not have scoliosis and/or kyphosis.
- The beneficiary does not have an inpatient admission for Spinal Fusion Except Cervical within the 120 days prior to the episode.
- The beneficiary does not have a diagnosis of fracture on the trigger claim (i.e., procedure was not due to trauma).
- The episode does not have IP procedures billed without relevant MS-DRG codes.
- The episode can be attributed to at least one main clinician.
- The episode trigger claim was in an acute IP hospital, OP hospital, HOPD, ambulatory/office-based care, or ASC setting.
- The episode is not an outlier case.

To determine whether the Lumbar Fusion measure's inclusion criteria distort patient characteristics on episodes, we produced and analyzed distributions of patient characteristics (age, race, sex, dual eligibility status, income, unemployment, hierarchical condition categories [HCCs]) for (i) episodes with inclusion criteria, (ii) episodes without inclusion criteria, (iii) beneficiaries with inclusion criteria, and (iv) beneficiaries without inclusion criteria.

This analysis shows that the Lumbar Fusion measure's inclusion criteria have a small effect on the percentage of beneficiaries of any particular demographic or patient characteristic. The difference between beneficiaries being included or not included in the measure is less than 1.8 percentage points across each of the characteristics in the analysis at TIN level testing, and less than 2.1 percentage points at TIN-NPI level testing. To illustrate, the percent of beneficiaries aged 65 to 69 without applying the inclusion criteria is 28.3 percent, compared to 28.5 percent at both TIN and TIN-NPI level testing. The difference in the percentage of beneficiaries for race with and without the inclusion criteria is less than 0.7 percentage points for each of the race categories, with the exception of white, where the difference is within 1.1 and 1.5 percentage points at TIN and TIN-NPI level testing, respectively, when inclusion criteria are applied. The share of male and female beneficiaries remains the same when comparing the use of inclusion criteria at TIN-NPI level testing, with 57.9 percent female and 42.1 percent male either with or without the application of inclusion criteria. At TIN level testing, there is a difference of 0.1 percentage points between the share of male and female beneficiaries. These results indicate that there is minimal shift in patient characteristics as a result of using the inclusion criteria listed above at both TIN and TIN-NPI level testing.

3.1.7 Sample Differences

n/a

3.1.8 Social Risk Factors Included in Analysis

The social risk factors analyzed were variables from the ACS, EDB, and CME. All ACS variables are at the Census Block Group level. Social risk variables analyzed include the following:

- 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
- 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
- Employment (ACS)
 - Unemployment Rate > 10%
 - Unemployment Rate <= 10%
- Race (EDB)
 - Asian, Black, Hispanic, North American Native, White, and Other
- Sex (EDB)
 - Female, male
- Dual status (CME)
 - Full dual, partial dual, non-dual

3.2 Reliability Testing

3.2.1 Level of Reliability Testing

The following levels of reliability were tested: critical data elements used in the measure and performance measure score (e.g., signal-to-noise analysis).

3.2.2 Method of Reliability Testing

Data Element Reliability

The Lumbar Fusion measure is constructed using CMS claims data, as described in Section 3.1.2., CMS has implemented several auditing programs to assess overall claims code accuracy, ensure appropriate billing, and 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, and formerly Program Safeguard Contractors, to ensure program integrity; the agency also uses Recovery Audit Contractors 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 2017, CERT estimates that proper payment, which includes payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments

each year.¹⁷ The Fiscal Year (FY) 2018 Medicare FFS program proper payment rate was 91.9 percent.¹⁸ CMS continues to perform successful corrective actions and give providers additional education to ensure accurate billing.

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

Measure Reliability

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 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.

This approach 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

σ_b^2 is the between-group variance of clinicians within the episode group

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 the systematic differences between the clinician and their peer cohort.

3.2.3 Statistical Results from Reliability Testing

Measure Reliability

At the 10, 20, and 30-episode thresholds, 100 percent of TINs and TIN-NPIs have a mean reliability greater than or equal to 0.4. At a volume threshold of at least 10 episodes, the mean reliability is 0.77 for TINs and 0.69 for TIN-NPIs. The mean reliability continues to increase at the 20 and 30-episode volume thresholds.

Table 2: Reliability Results at Various Volume Thresholds

Volume Threshold (# episodes)	TIN		TIN-NPI	
	Mean Reliability	% ≥ 0.4	Mean Reliability	% ≥ 0.4
10	0.77	100.0%	0.69	100.0%
20	0.84	100.0%	0.79	100.0%
30	0.88	100.0%	0.84	100.0%

¹⁷ Comprehensive Error Rate Testing (CERT) Program. "Appendices Medicare Fee-for-Service 2018 Improper Payments Report". Table A6. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicare-FFS-Compliance-Programs/CERT/Downloads/2018MedicareFFSSupplementalImproperPaymentData.pdf>

¹⁸ Ibid.

3.2.4 Interpretation

Measure Reliability

Overall reliability of the Lumbar Fusion measure is very high at a volume threshold of 10 episodes or more for TINs and exceeds 0.4 for TIN-NPIs, due to the large number of episodes attributed to clinicians. CMS generally considers 0.4 as the threshold indicating ‘moderate’ reliability, which is supported by previous work into reliability.¹⁹

While higher volume thresholds yield even higher reliability results, it is at the cost of further reducing the number of clinicians and clinician groups able to receive a measure score.

3.3 Validity Testing

3.3.1 Level of Validity Testing

We conducted performance measure score validity testing, which included systematic assessment of face validity and empirical validity testing.

3.3.2 Method of Validity Testing

Face Validity

The Lumbar Fusion measure was developed through a structured, iterative process for gathering detailed input from recognized clinician experts on the measure. These expert panels were convened to methodically assess the extent to which the measure: (i) captured what it was intended to capture, and (ii) differentiated between provider performance. Experts in this clinical area evaluated specifications in an iterative process to ensure that each aspect of the measure (e.g., assigned services) was intentionally capturing only the costs of care within the reasonable influence of the attributed clinician for a defined patient population (i.e., the ability of the measure score to differentiate good from poor performance).

In developing and refining this measure, 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 which measure to develop, on the measure scope, and on the composition of a smaller, targeted workgroup to provide detailed input on each aspect of measure specifications. The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels workgroup was composed of 13 members, affiliated with 13 specialty societies, including 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 a patient and family perspective. PFC input included concepts of healthcare quality and value, guiding principles and measure-specific input to inform the workgroups such as pre- and post-

¹⁹ Mathematica, Inc., “Memorandum: Reporting Period and Reliability of AHRQ, CMS 30-Day and HAC Quality Measures – Revised,” http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability-.pdf.

trigger windows for selected episodes, and inclusion of services and costs for attributed clinicians. 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.

One of the key roles of the measure-specific workgroup was to develop service assignment rules for the cost measure. These service assignment rules are intended to ensure clinicians are evaluated on services and costs that are clinically related to the attributed clinician's role in lumbar fusion surgery, thus preventing inclusion of unrelated cost variation in this measure. Assigned services occurring in the emergency department, OP facility and clinician services, IP – medical, IP – surgical, IRF – medical, DME, and HH setting were defined separately for the pre- and post-trigger windows, and include lumbar fusion, evaluation, testing, treatment, complications, and follow-up.

Empirical Validity Testing

We undertook two approaches to estimate the measure's validity. In the first approach, we evaluated the empirical validity of the Lumbar Fusion measure by examining differences in risk adjusted cost for known indicators of resource or service utilization based on a literature review, specifically complications related to lumbar fusion. For this analysis, we compared the ratio of observed to expected cost (henceforth called the "O/E cost ratio") for Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episodes with and without complications related to lumbar fusions that occur in the post-trigger period. This analysis sought to confirm the expectation that the Lumbar Fusion measure captures variation in service utilization.

In the second approach, we evaluated how different types of cost impact risk-adjusted measure scores. Certain services or costs included in the Lumbar Fusion measure were classified into clinically coherent groups of services, called "clinical themes." The Lumbar Fusion measure clinical themes are:

- **Preoperative Work-Up:** Includes routine chest x-rays, electrocardiogram, and laboratory testing, such as blood tests to assess coagulation; other diagnostic techniques, such as x-rays of the spine; or diagnostic procedures, such as office or outpatient evaluations.
- **Anesthesia/Pain Management:** Includes IP and OP hospital care including emergency department visits or critical care provided for poisoning by opioids or adverse effects of other synthetic narcotics, muscle spasms, or pain; or imaging, testing, medications, catheters, and other diagnostic and therapeutic procedures and related supplies.
- **Wound Care:** Includes removal of sutures, change or removal of drains, aftercare following the surgery, care for a post-procedural hematoma or seroma, and care to treat complications of the surgical wound.
- **Post-Acute Care:** Includes IP or critical care for follow up on the completed treatment, and any SNF, IRF, and HH care related to complications from the procedure.
- **Durable Medical Equipment (DME):** Includes DME and supplies, such as walkers, wheelchairs, canes, tape, catheters, drainage bags, and wound care required post surgery.
- **Cardiovascular Complications:** Includes IP and OP hospital care including emergency department visits or critical care related to respiratory arrest, pericardial effusion, atrial fibrillation or flutter, or other arrhythmias; imaging including CT scans or echocardiograms; or diagnostic and therapeutic procedures, blood transfusion,

medications, related supplies and percutaneous cardiovascular procedures within the appropriate time frame.

- **Thromboembolism (DVT/PE):** Includes IP and OP hospital care including emergency department visits and critical care for embolisms and thromboses, including diagnostic or therapeutic procedures to diagnose and treat these conditions, such as laboratory tests, imaging, medications, and supplies.
- **Infection and GI Complications:** Includes IP and OP hospital care including emergency department visits and critical care for paralytic ileus or intestinal obstructions, pain, intestinal disorders, pyelonephritis, or other complications, including diagnostic or therapeutic procedures to diagnose and treat, such as tests, imaging, medications, and supplies.
- **Mechanical Complication / Need for Revision:** Includes IP and OP hospital care including emergency department visits and critical care for spinal stenosis, disc displacement or degeneration, or other complications, including diagnostic or therapeutic procedures, imaging, spinal fusion, medications and related supplies.
- **Neurological Complications:** Includes IP and OP hospital care including emergency department visits and critical care for anesthesia or paresthesia of skin, other disturbances of skin sensation, muscle weakness, or altered mental status, including diagnostic or therapeutic procedures, imaging, medications and related supplies.

As with the first analysis for validity, the aim of this analysis was to determine whether the measure is capturing variation in provider cost in the manner intended and expected. To measure this, we calculated the Pearson correlation between the cost of each clinical theme and the overall risk-adjusted cost for an episode.

We expected that the Mechanical Complications / Need for Revision theme would have the highest correlation with risk-adjusted episode cost, as complications are likely associated with high cost even after accounting for beneficiary characteristics.²⁰ We would expect similar trends for the Neurological Complications and Cardiovascular Complications themes as they contain services relating to complications, such as altered mental status and pericardial effusion. By contrast, we expected that Preoperative Workup and DME themes will likely have variations in cost largely due to beneficiary characteristics. As a result, the correlation between the cost of these themes and the overall risk-adjusted cost will be much lower than for complications.

3.3.3 Statistical Results from Validity Testing

Table 3 presents an analysis of validity, showing the O/E cost ratio of episodes with or without downstream acute (re)admissions and with or without Post-Acute Care. The mean O/E cost ratio for all episodes is 1.00. The mean O/E cost ratio for episodes with downstream acute (re)admission during the post-trigger period is 1.48, compared with 0.99 for episodes without downstream acute (re)admission during the post-trigger period. The mean O/E cost ratio for episodes with post-acute care during the post-trigger period is 1.12, compared with 0.90 for episodes without post-acute care during the post-trigger period.

²⁰ Khan, N.A., Quan, H., Bugar, J.M. et al., "Association of postoperative complications with hospital costs and length of stay in a tertiary care center" J Gen Intern Med (2006) 21: 177.

Table 3: Distribution of Observed to Expected Ratios

Episode Type	Observed / Expected Ratio										
	Mean	Std. Dev.	Percentile								
			1st	5th	10th	25th	50th	75th	90th	95th	99th
All Final Episodes	1.00	0.26	0.39	0.74	0.79	0.86	0.93	1.06	1.34	1.52	1.91
Episodes with Downstream Acute (Re)admission	1.48	0.52	0.72	0.88	0.96	1.08	1.36	1.76	2.13	2.50	3.25
Episodes without Downstream Acute (Re)admission	0.99	0.24	0.38	0.74	0.79	0.86	0.93	1.05	1.30	1.48	1.79
Episodes with Post-Acute Care (IRF LTCH HH SN)	1.12	0.28	0.67	0.79	0.85	0.94	1.04	1.25	1.50	1.65	1.98
Episodes without Post-Acute Care (IRF LTCH HH SN)	0.90	0.19	0.27	0.72	0.77	0.83	0.89	0.94	1.00	1.10	1.71

The clinical themes analysis demonstrates that there is a strong correlation between the Post-Acute Care (correlation: 0.70) theme and risk-adjusted cost. The Mechanical Complication / Need for Revision theme has the second strongest correlation (correlation: 0.52) with the risk-adjusted cost. By contrast, the Durable Medical Equipment (DME) (correlation: -0.02) and Preoperative Work-Up (correlation: 0.03) themes had much lower correlation with risk-adjusted cost. The negative correlation for the Durable Medical Equipment (DME) theme suggests that costs are slightly lower than the risk adjustment model predicts.

3.3.4 Interpretation

As expected, the average O/E cost ratio for episodes with post-trigger complications is higher than for episodes without downstream complications. This result demonstrates that the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure is able to accurately capture higher resource use.

The clinical themes analysis demonstrates that high risk-adjusted cost is associated with themes related to complications, with 3 of the 4 highest correlations being Infection and GI Complications, Neurological Complications, and Mechanical Complication. Post-Acute Care is also a large opportunity to capture the care for complications related to the procedure. This indicates that the measure may penalize clinicians who have higher rates of complications, while not disincentivizing the provision of appropriate pre- and post-operative care, such as preoperative work ups and wound care.

3.4 Exclusions Analysis

3.4.1 Method of Testing Exclusions

Exclusions are used in Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels either to capture a homogenous patient population within the scope of the measure focus on lumbar fusion and that episodes provide meaningful information to attributed clinicians or as part of data processing to ensure that sufficient data are available to accurately determine episode spending and calculate risk adjustment for each episode. For the exclusions analysis, we focused on exclusions added to ensure a homogenous patient population. These exclusions, along with their rationales, are listed below:

- *Episodes where beneficiary death date occurred before the episode end date.*

- These episodes were excluded for all measures due to the potential to reflect inaccurately a clinician's performance. Episodes where the beneficiary died may be unusually high-cost, due to perimortem treatment costs, or unusually low-cost, due to the truncated episode window. Neither of these cases accurately reflects the efficiency of the clinician performing the treatment.
- *Episodes where the beneficiary has cancer.*
 - Beneficiaries with specific malignant neoplasms are excluded from this measure, as defined by the presence of ICD-10 diagnosis codes on the Part B Physician/Supplier trigger claims. Cancer patients likely require more complex surgery and have more complex post-surgical care needs than patients with degenerative spinal conditions.
- *Episodes where the beneficiary has an osteoporotic compression fracture.*
 - Beneficiaries with osteoporotic compression fractures are excluded from this measure, as defined by the presence of ICD-10 diagnosis codes on the Part B Physician/Supplier trigger claims. Patients with osteoporotic compression fractures likely require more complex surgery and may have more difficulty with healing than patients with degenerative spinal conditions.
- *Episodes where the beneficiary has an infection.*
 - Beneficiaries with infections are excluded from this measure, as defined by the presence of ICD-10 diagnosis codes on the Part B Physician/Supplier trigger claims. Patients with infections may require more complex surgery, require antibiotic treatment, and may have more complex post-surgical care needs than patients with degenerative spinal conditions.
- *Episodes where the beneficiary is undergoing a redo lumbar fusion.*
 - Beneficiaries undergoing a redo lumbar fusion are excluded from this measure, as defined by ICD-10 and CPT/HCPCS codes on the Part B Physician/Supplier claims, IP claims, and OP claims on the procedure date. Patients with redo lumbar fusions may require more complex surgical procedures and likely require more complex post-care including pain management than patients after initial spinal fusions.
- *Episodes where the beneficiary has experienced trauma due to a fracture.*
 - Beneficiaries that have experienced trauma from a fracture are excluded from this measure, as defined by the presence of an ICD-10 diagnosis code for a fracture on the Part B Physician/Supplier trigger claims. Trauma patients may have different and more severe injuries to nerves and surrounding structures, may require more extensive surgeries, and are likely to have more complex post-surgical care.
- *Episodes where the beneficiary has scoliosis and/or kyphosis.*
 - Patients with scoliosis and kyphosis often require different fusion techniques and likely require more complex post-surgical care.
- *Episodes where a Spinal Fusion Except Cervical occurs within 120 days prior to the episode.*
 - Episodes where there is a Spinal Fusion Except Cervical within 120 days prior to the episode is excluded from this measure, as defined by the presence of an MS-DRG on IP claims during the 120 lookback period prior to the trigger. A patient who had a prior spinal fusion within the past 120 days will likely require more complex post-surgical care so will differ from the type of patient in this episode.
- *Episodes where a lumbar fusion with curvature, malignancy, infections, or extensive fusions occurs.*

- Episodes where any lumbar fusions with curvature, malignancy, infections, or extensive fusions occur are excluded from this measure, as defined by the presence of MS-DRGs 456-458 on IP claims during the trigger event. The cases that are grouped into MS-DRGs 456-458 typically have more complex surgery and more complex post-surgical care so differ from the cases in this episode.
- *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.

Given the rationales for these exclusions, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a substantially higher or lower mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). For the exclusions, we examined the number of episodes and beneficiaries affected, as well as the distributions of observed cost and O/E cost ratio (calculated by applying existing risk factor coefficients to the excluded episodes) for excluded episodes. We then compared the cost characteristics of the excluded episodes to those of final episodes included in measure calculation to assess the distinctness between the two patient cohorts. A full list of the exclusions and details used for the Lumbar Fusion measure is provided in the Measure Codes List.²¹

3.4.2 Statistical Results from Testing Exclusions

Table 4 below presents observed cost statistics and O/E cost ratios for the Lumbar Fusion measure exclusions. Cost statistics are also provided for the set of final episodes included in the Lumbar Fusion measure for comparison, with a testing volume threshold of 10 episodes at the TIN and TIN-NPI levels.

Table 4: Cost Statistics for Measure Exclusions

Exclusion	Episodes		Observed Cost			O/E		
	#	%	Mean	Percentile		Mean	Percentile	
				10 th	90 th		10 th	90 th
All Episodes Meeting Triggering Logic	79,067	100.00%	\$37,504	\$22,717	\$59,784	0.99	0.76	1.33
Cancer	215	0.27%	\$20,158	\$3,711	\$59,970	0.76	0.21	1.07
Curve Cancer Infection	9,052	11.45%	\$39,278	\$5,204	\$78,057	0.93	0.33	1.31
Death in Episode	782	0.99%	\$41,187	\$4,838	\$75,360	1.00	0.32	1.41
Infection	657	0.83%	\$38,880	\$4,365	\$88,648	0.88	0.21	1.35
Osteoporotic Compression Fracture	258	0.33%	\$35,369	\$3,904	\$69,780	0.87	0.21	1.30
Outlier Cases	1,084	1.37%	\$66,106	\$23,446	\$117,552	1.51	0.53	2.70
Redo Lumbar Spine Fusion	7,595	9.61%	\$39,928	\$24,418	\$65,967	0.99	0.74	1.31
Scoliosis/Kyphosis	7,991	10.11%	\$41,263	\$6,247	\$82,670	0.97	0.46	1.32
Spinal Fusion Except Cervical	98	0.12%	\$33,763	\$4,157	\$67,546	0.93	0.18	1.32
Trauma	1,627	2.06%	\$37,302	\$4,947	\$71,578	0.93	0.33	1.31
<i>Final Episodes (TIN)</i>	48,870	61.81%	\$35,970	\$27,184	\$53,975	0.99	0.79	1.32
<i>Final Episodes (TIN-NPI)</i>	41,622	52.64%	\$35,905	\$27,168	\$53,860	0.99	0.78	1.30

²¹ CMS, "Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure Codes List," *MACRA Feedback Page*, <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2019-revised-ebcm-measure-specs.zip>.

3.4.3 Interpretation

The statistical results indicate that the majority of excluded episodes differ substantially in the mean observed cost and mean O/E cost ratio and that they have larger variation compared to the final set of episodes. These results support the exclusion of these episodes to ensure a comparable patient cohort that will yield meaningful information to attributed clinicians. Further discussion of the results for each exclusion is provided below.

Episodes ending in death: The mean episode cost for episodes ending in death is approximately \$5,000 more than the final set of episodes, due to potentially high costs of services related to life saving services or underlying expensive comorbidities unrelated to the procedure. The difference becomes more pronounced at the 90th percentile, where episodes ending in death are more than \$20,000 more than the final set of episodes at the TIN and TIN-NPI levels. The measure seeks to avoid problematic incentives that could lead to risks of clinicians cherry picking to avoid high-risk patients.

Episodes where the beneficiary has cancer: While the mean observed cost is about \$15,000 lower for these episodes than the final set of episodes included in the measure, this small number of episodes with specific malignant neoplasm diagnoses are excluded as the care pathways for these patients are different, and the focus of the measure is on degenerative spinal conditions rather than spinal conditions that are a result of cancer.

Episodes where a lumbar fusion with curvature, malignancy, infections, or extensive fusions occurs: The mean episode cost is approximately \$4,000 more than the final set of episodes and are typically reflective of more complex surgery and post-surgical care, outside the scope of the measure. There is also a high level of variation in the episodes cost, with a much lower observed cost at the 10th percentile (\$5,204) compared to a much higher observed cost at the 90th percentile (\$78,057) compared to the final episodes. This high level of variation also suggests it is appropriate to exclude these episodes.

Episodes where the beneficiary has an infection: The mean episode cost is approximately \$3,000 more than the final set of episodes, and although it is only slightly more expensive, it has the potential to become very expensive at the higher percentiles, almost \$35,000 more than the final set of episodes at the 90th percentile. Due to the potential for more complex surgery, the requirement of antibiotic treatment, and more complex post-surgical care it does not fit within the patient cohort the measure is designed for, those with degenerative spinal conditions.

Episodes where the beneficiary has an osteoporotic compression fracture: Although the mean episode cost is approximately the same as the final set of episodes, episodes in the right tail are substantially more expensive, about \$14,000 more than the final set of episodes. Patients with osteoporotic compression fractures often have more difficulty healing than patients with degenerative spinal conditions, and are considered outside the scope of the measure's intended patient cohort.

Episodes where the beneficiary is undergoing a redo lumbar fusion: The mean episode cost is approximately \$4,000 more than the final set of episodes. This is excluded because they likely require more complex procedures and post-care than patients after initial spinal fusions.

Episodes where the beneficiary has scoliosis and/or kyphosis: The mean episode cost is approximately \$5,000 more than the final set of episodes. Scoliosis and Kyphosis patients composes are within a different cohort as they often require different fusion techniques, so this would reflect differently from the standard lumbar fusion the measure aims to capture.

Episodes where a Spinal Fusion except Cervical occurs during the lookback period: There are only 98 episodes in this exclusion category, and patients who have already received prior spinal fusion in the past 120 days may require more complex surgical care resulting in particularly high-cost episodes in the right tail of the distribution (approximately \$13,000 more than the final set of episodes).

Episodes where the beneficiary has experienced trauma: The mean episode observed cost is approximately \$2,000 more than the final set of episodes and has been excluded as these reflect a different patient cohort with different or more severe injuries to nerves and surrounding structures, and may require more extensive surgeries than is intended to be captured in this lumbar fusion measure that focuses on degenerative disease.

Episodes classified as outlier cases: The O/E cost ratio for outlier cases ranges from 0.53 at the 10th percentile to 2.70 at the 90th percentile, indicating that the risk adjustment model is currently unable to account for the patient characteristics associated with these high- and low-cost outlier episodes. Excluding outliers based on risk-adjusted cost eliminates the episodes that deviate most from expected spending levels based on patient characteristics.

3.5 Risk Adjustment or Stratification

3.5.1 Method of Controlling for Differences

Differences in case mix are controlled for using a statistical risk model with 122 risk factors and stratification by three risk categories.

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 beneficiary's Parts A and B claims during the period 120 days prior to the episode trigger and are specified in the CMS-HCC Version 22 (V22) 2016 model. Episodes for beneficiaries without a full 120-day lookback period are excluded from the measure. This 120-day period is used to measure beneficiary health status and ensures that each beneficiary's claims record contains sufficient fee-for-service data both for measuring spending levels and for risk adjustment purposes.

In addition, the risk adjustment model includes status indicator variables for whether the beneficiary qualifies for Medicare through Disability or ESRD. The model also includes an indicator of whether the beneficiary recently required long-term care, defined as 90 days in a long-term care facility without being discharged to community for 14 days. Beneficiaries 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 costs that attributed clinicians can reasonably influence, informed by expert clinician input and empirical analyses. The following variables were added to avoid potential unintended consequences:

- whether the procedure was an anterior interbody fusion, which involves a more invasive surgical approach so patients may have more complex post-operative care,
- whether the procedure was as same-day anterior and posterior lumbar fusion because anterior and posterior fusion performed on the same day involves a more invasive surgical approach so patients may have more complex post-operative care,
- whether the patient has a history of or current use of anticoagulants because they will likely require more post-surgical monitoring for the conditions that led to anticoagulant therapy,
- whether the procedure was a combined posterior or posterolateral and posterior interbody fusion because the combination of the two approaches involve more complex surgery and differing post-surgical care,
- whether the patient has hypertension, which have a higher risk of cardiovascular complications from the surgery and could have higher costs outside of the clinician's influence,
- whether the patient has morbid obesity or obesity, which confers a much higher risk of pulmonary, metabolic and cardiovascular complications from the surgery and could have higher costs outside the clinician's influence,
- whether the patient has osteoporosis, as they will be at a higher risk during surgery and may require different approaches and management outside the influence of the clinician,
- whether the procedure was a posterior or posterolateral fusion as this requires differing post-surgical care,
- whether the procedure was a posterior interbody fusion, as this requires differing post-surgical care,
- whether the 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,
- whether the patient smokes, as smoking confers a higher risk of pulmonary and cardiovascular complications from the surgery,
- place of setting for acute IP hospitals, HOPD, ASC, 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,
- whether the patient has frailty indicators (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;
- whether the patient experienced a recent hospitalization for medical back problems within 120 days of the trigger, as hospitalization for back problems indicates a more severe condition.

As with the CMS-HCC model, the risk adjustment approach for this measure uses an ordinary least squares 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 cost 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, as noted in the exclusions analysis above, 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 ensure that average expected costs are the same after outlier removal.

Finally, the risk adjustment model outlined above is performed separately for each of the four Lumbar Fusion measure sub-groups, which are based on the level of fusion, are:

- One-level Lumbar Fusion
- Two-level Lumbar Fusion
- Three-level Lumbar Fusion

Full details of the risk adjustment model are in the Measure Codes List File.²² The National Summary Data Report (NSDR) Addendum includes regression coefficients and standard errors for each of the covariates used in the risk adjustment model.²³

3.5.2 Conceptual, Clinical, and Statistical Methods

We selected the CMS-HCC model based on previous studies evaluating its appropriateness for use in risk adjusting Medicare claims data. This 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 beneficiaries. In addition, 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. Because the CMS-HCC model has already been extensively tested, we focus our testing on how the CMS-HCC model was adapted to the Lumbar Fusion measure methodology.

The workgroup provided input on measure-specific risk adjusters after reviewing empirical analyses on subpopulations of interest to assess whether and if so, how, particular factors should be accounted for in the model. These could include patient characteristics, factors outside the influence of the attributed clinician, or any other factors that would help prevent unintended consequences. These additional risk adjusters are listed in the section above.

As previously noted, the risk adjustment model is run on episodes stratified into sub-groups, which may qualify as "ordering" of risk factors. Sub-groups were also determined based the workgroup's input, with the goal of ensuring clinical comparability among episodes so that the cost measure fairly compares clinicians with similar patient case-mix. The sub-groups, which are based on level of fusion, are listed in the above section. The sub-groups were developed because the single level fusion patients have simpler surgery and may have a different recovery pattern than patients with more levels of fusions, such as two-level and three-level fusions, which progressively become more complex.

Information on data sources and methodology used to analyze social risk factors can be seen in Section 3.1.8.

²² CMS, "Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure Codes List," *MACRA Feedback Page*, <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2019-revised-ebcm-measure-specs.zip>.

²³ CMS, "National Summary Data Report Addendum: 11 Episode-Based Cost Measures and Revised MSPB Clinician Measure," *MACRA Feedback Page*, <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/MACRA-Feedback.html>.

3.5.3 Conceptual Model of Impact of Social Risks

Our conceptual model of the impact of social risk factors is informed by both published, peer-reviewed literature and data analysis.

3.5.4 Statistical Results

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 NQF #2158: MSPB-Hospital cost measure). Recalling that the risk model relies on the existing CMS-HCC model, testing results for factors included in the CMS-HCC V22 2016 model can be found in the Pope et al (2011) report.²⁴ For measure-specific factors not included in the CMS-HCC model, we sought expert clinician input through the workgroup, which provided recommendations on additional risk adjusters and sub-groups.

The results of the statistical analysis used to characterize our risk adjustment model can be found in the NSDR Addendum, which includes regression coefficients and standard errors for each of the covariates used in the risk adjustment model.

3.5.5 Analyses and Interpretation in Selection of Social Risk Factors

Acumen analyzed gender, dual status, income, education, and unemployment as social risk factors (more information on these variables can be found in Section 3.1.8). Beneficiary gender and dual status were obtained from the EDB and CME. Information on income, education, and unemployment was obtained from ACS data and linked to episodes by census block group where possible to provide a more granular level of analysis than ZIP code.

The percentage of female beneficiaries range from 56.2 to 58.9 percent across the three sub-groups in this measure. The majority of the beneficiaries (87.0 percent - 87.4 percent) have non-dual status. Income level is categorized into high, medium, and low from the continuous average income variable in ACS; therefore, each category has 33 percent of observations. Less than 2.1 percent of beneficiaries are classified below a high school education, and approximately 98 percent of episodes are classified at a high school level or greater across the three sub-groups. Finally, 21.1 to 23.4 percent of beneficiaries have high unemployment designation (>10%).

Acumen examined the impact of including social risk factors into 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 standard set of risk adjustment variables from the CMS-HCC V22 2016 model, disability status, ESRD status, interaction variables, recent long-term care use, and measure-specific clinical risk adjusters. Acumen ran a step-wise regression to include gender, dual status, gender + dual status, and gender + dual + income + education + unemployment + race, on top of the adapted CMS-HCC model. The step-wise regressions help evaluate individual as well as joint significance of the social risk factors. We examined the impact of including social risk factors into our risk adjustment model with T-test of individual significance and F-test of joint significance.

First, we analyzed the model coefficients and p-values for each of the base and social risk factor models to understand whether any of the social risk factor covariates are predictive of episode cost. The T-test and F-test revealed many significant p-values, indicating that social risk factors are likely predictive factors for determining resource use among beneficiaries for the relevant

²⁴ 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.

characteristic. However, the analysis also shows that the directions of the effects of social risk factors is not consistent.

Secondly, we analyzed the impact of adding social risk variables on overall model performance by looking at the differences in the O/E cost ratios with and without social factors in the risk adjustment model. When including social risk factors in our risk adjustment regression, the minor differences in the O/E cost ratios, even for providers at high or low extremes of risk, indicates that social risk factor effects on the model performance are likely captured through existing risk adjustment variables. When including the social risk factors in risk adjustment, the O/E cost ratios changed by ± 0.03 or less for 98.1 percent of TINs and 98.1 percent of TIN-NPIs.

Finally, we analyzed the correlation between measure scores calculated with and without the social risk factors. The measure scores calculated with and without these social factors were highly correlated at both the TIN and TIN-NPI levels, with a Spearman correlation coefficient of 0.994 and 0.993, respectively. These results indicate that the inclusion of social risk factors in the current risk adjustment model would have a limited effect on measure scores.

Due to the inconsistent direction and limited impact of social risk factor effects under the current risk adjustment model, we believe the Lumbar Fusion measure's risk adjustment model sufficiently accounts for the effects of social risk factor on clinician measure scores.

3.5.6 Method for Statistical Model or Stratification Development

To analyze the validity of current risk adjustment model, we examined three analyses: (1) R-squared and adjusted R-squared for the regression models, (2) O/E cost ratios and predictive ratios to examine the fit of the models at different levels of patient complexity, and (3) coefficient estimates, standard errors, and p-values for each sub-group.

- 1) *R-squared and adjusted R-squared* were calculated for the measure overall as well as for each sub-group. The results should be evaluated in the context of the service assignment rules, which indicate which costs are counted in the measures and which costs are not counted. This is an important distinction from all-cost measures, as a low R-squared does not necessarily indicate that a measure reflects variation unrelated to clinical care, while a high R-squared does not necessarily indicate the opposite; instead, the risk adjustment models must be evaluated in concert with the service assignment rules. These results are provided in Section 3.5.7.
- 2) *Predictive ratios and O/E cost ratios* were calculated for each "risk decile" for the episode group. A "risk decile" is based on the risk scores, which indicate how costly episodes are expected to be, as predicted through risk adjustment. After arranging episodes into deciles based on their risk score, we calculated the predictive ratios and average O/E cost ratios for each decile. The predictive ratio aims 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, and is calculated using the formula of average (expected cost)/average (observed cost) for all episodes in each decile. Similarly, the O/E cost ratio demonstrates the model's prediction accuracy, and is calculated using the formula of average (observed cost/expected cost) for all episodes in each decile. These are discussed in Sections 3.5.8 and 3.5.9.
- 3) *Coefficient estimates, standard errors, and p-values* were run for each sub-group 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 and set of sub-groups, 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.

The results of these analyses are presented in the NSDR Addendum to aid in the overall assessment of the predictive ability of the risk adjustment models.²⁵

3.5.7 Statistical Risk Model Discrimination Statistics

The overall R-squared for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure, calculated by dividing explained sum of squares by total sum of squares is 0.50. The adjusted R-squared is also 0.50.

The NSDR Addendum 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. 2011.²⁶

3.5.8 Statistical Risk Model Calibration Statistics

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. The average O/E cost ratio is 1.01 for most deciles except for Decile 2, with 1.02, indicating that the model is accurately predicting actual episode cost. Full results are presented in the NSDR Addendum.

3.5.9 Statistical Risk Model Calibration – Risk Decile

Analysis of predictive ratios by risk decile for the measure shows that the model has consistent predictive ratios across risk score deciles, with each decile having a predictive ratio of 1.00 or 0.99.

3.5.10 Results of Risk Stratification Analysis

Results indicate that the three measure sub-groups have varying measure scores (see below table). Specifically, One-level Lumbar Fusions are less expensive than Two-level and Three-level Fusions. At the TIN level, the mean score for One-level Lumbar Fusion episodes is \$33,891 compared to Two-level Lumbar Fusion episodes at \$39,395 and Three-level Fusion episodes at \$44,387. At the TIN-NPI level, there is the same trend of increased cost with increased level of fusions. The mean score for One-level Lumbar Fusion episodes is \$33,750 compared to Two-level Lumbar Fusion episodes at \$39,441 and Three-level Fusion episodes at \$44,287. Thus, the three levels of fusions are treated separately due to their varied costs. The level of fusion required for the procedure is related to the beneficiaries' underlying health conditions. Single level fusion patients have simpler surgery and may have different recovery patterns than patients with more level of fusions. A similar distinction exists between two-level and three-level fusions. Stratifying episodes into these sub-groups helps ensure meaningful comparison of clinician resource use.

Table 5: Distribution of Score by Sub-Group

Level	Sub-group	Provider Count	Mean Score	Score Percentile						
				1st	10th	25th	50th	75th	90th	99th
TIN	All TINs	1,440	\$36,631	\$29,418	\$32,878	\$34,478	\$36,282	\$38,427	\$41,033	\$46,349

²⁵ CMS, "National Summary Data Report Addendum: 11 Episode-Based Cost Measures and Revised MSPB Clinician Measure," *MACRA Feedback Page*, <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/MACRA-Feedback.html>.

²⁶ 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.

Level	Sub-group	Provider Count	Mean Score	Score Percentile						
				1st	10th	25th	50th	75th	90th	99th
TIN	One-level Lumbar Fusion	1,436	\$33,891	\$27,365	\$30,465	\$31,702	\$33,346	\$35,463	\$37,990	\$45,947
TIN	Two-level Lumbar Fusion	1,417	\$39,395	\$28,783	\$33,426	\$35,788	\$38,933	\$42,258	\$46,191	\$57,004
TIN	Three-level Lumbar Fusion	1,071	\$44,387	\$27,903	\$34,916	\$38,750	\$43,126	\$48,461	\$55,814	\$70,105
TIN-NPI	All TIN-NPIs	3,286	\$36,537	\$29,697	\$32,630	\$34,156	\$36,062	\$38,317	\$41,316	\$47,466
TIN-NPI	One-level Lumbar Fusion	3,266	\$33,750	\$26,934	\$30,156	\$31,426	\$33,132	\$35,384	\$38,083	\$46,049
TIN-NPI	Two-level Lumbar Fusion	3,190	\$39,441	\$28,727	\$32,863	\$35,267	\$38,538	\$42,494	\$47,378	\$59,783
TIN-NPI	Three-level Lumbar Fusion	2,175	\$44,287	\$27,577	\$33,910	\$37,621	\$42,238	\$48,841	\$57,271	\$78,462

3.5.11 Interpretation

The R-squared values for the model, which measure the percentage of variation in results predicted by the model, are higher than the values presented in similar analyses of risk adjustment models.²⁷ As noted in Section 3.5.6, these results should be interpreted alongside service assignment rules, which remove clinically unrelated services, so the resulting variation is reflective of variation related to factors within a clinician's reasonable influence.

As demonstrated in Section 3.5.8 and 3.5.9, the average O/E cost ratios and the predictive ratios for all risk deciles are close to one. Predictive ratios close to one indicate that expected spending is accurately predicting observed spending. Overall, the results show that the model is accurately predicting observed spending, regardless of overall risk level.

3.6 Identification of Meaningful Differences in Performance

3.6.1 Method

Our method of determining clinically meaningful differences in episode-based cost measure scores consists of stratifying the clinician measure scores by meaningful characteristics and investigating the clinician score distribution by percentile. Stratification is performed for each of the following characteristics: urban/rural, census division, census region, risk score, and the number of episodes attributed to the clinician. We analyze the distribution of measure scores for clinicians defined by these characteristics, as well as for the overall episode group and for each sub-group.

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.

²⁷ 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

3.6.2 Statistical Results

Key findings show that, generally, there is a performance difference among clinicians in the Lumbar Fusion measure:

- (i) the 99th percentile of the measure score is 1.6 times the 1st percentile at the TIN-NPI level; and
- (ii) the 99th percentile of the measure score is nearly 1.8 times the 1st percentile at the TIN level.

These results indicate there is potential for saving Medicare spending.

The results also show that there is not systemic regional difference in clinician score. For instance, the mean scores for clinicians across nine census divisions (excluding 'Unknown') are within a less than \$2,100 range (i.e., \$35,815 - \$37,845 at the TIN level and \$35,749 - \$37,686 at the TIN-NPI level). Similarly, clinicians in urban areas seem to perform comparably to those in rural areas, with less than a \$700 difference.

In terms of other clinician characteristics, analysis of clinicians by number of episodes indicates that clinicians with more episodes perform similarly to those who perform fewer procedures. We also analyzed clinicians by risk score decile, as variation by risk score decile could indicate that the risk adjustment model is over- or under-correcting for clinicians with systematically riskier patients. Measure scores also show little variation by risk score decile, with a range in mean TIN score of \$35,592 to \$37,616 and a range in mean TIN-NPI score of \$35,888 to \$36,973, indicating that the risk adjustment model is overall functioning as intended. Full results can be seen in the NSDR.²⁸

3.6.3 Interpretation

There is clinically and practically significant variation in Lumbar 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. The measure-specific workgroup suggested development of sub-groups based on type of procedures, noting the differences in cost between One-level, Two-level, and Three-level Lumbar Fusion procedures. Overall, as expected, results show that clinicians are not being systematically penalized or rewarded due to risk score decile given the current Lumbar Fusion measure design (i.e., the differences in cost measure scores are not as a result of the risk profile of the patient cohort).

3.7 Missing Data Analysis and Minimizing Bias

3.7.1 Method

Since CMS uses Medicare claims data to calculate the Lumbar Fusion measure, Acumen expects a high degree of data completeness. To further ensure that we have complete and accurate data for each beneficiary who opens an episode, Acumen excludes episodes where beneficiary date of birth information (an input to the risk adjustment model) cannot be found in the EDB, the beneficiary does not appear in the EDB, or the beneficiary death date occurs before the episode trigger date.

The Lumbar Fusion measure also excludes episodes where the beneficiary is enrolled in Medicare Part C or has a primary payer other than Medicare in the 120-day lookback period and

²⁸ CMS, "National Summary Data Report: 11 Episode-Based Cost Measures and Two Revised Cost Measures, Updated Following Field Testing (Oct-Nov 2018)," *MACRA Feedback Page*, <https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/value-based-programs/macra-mips-and-apms/macra-feedback.html>.

episode window. In such situations, Medicare Parts A and B claims data may not capture the complete clinical profile for the beneficiary needed to capture the clinical risk of the beneficiary in risk adjustment. Furthermore, Parts A and B claims data may not capture all Medicare resource use if some portion of the beneficiary's care is covered under Medicare Part C.

3.7.2 Missing Data Analysis

The table below presents the frequency of missing data across the four categories of missing data that caused episodes to be excluded from the Lumbar Fusion measure. Frequency is presented in terms of the number of episodes excluded due to missing data, as well as the number of TINs and TIN-NPIs who had at least one episode excluded due to missing data. The missing data categories are:

- Beneficiary date of birth is missing
- Beneficiary death date occurred before the trigger date
- Beneficiary has a primary payer other than Medicare during the episode window or in the 120-day lookback period
- Beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 120-day lookback period and episode window

Table 6: Missing Data Categories for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure

Exclusion	# Episodes	# TINs	# TIN-NPIs
Missing birth date	0	0	0
Death before trigger	12	12	19
Other primary payer	11,089	2,372	6,785
Not continuously enrolled	4,380	1,738	4,286

3.7.3 Interpretation

As the Lumbar Fusion measure is calculated with Medicare claims data, Acumen expects a high degree of data completeness, which is supported by the limited frequency of missing data as noted above. Acumen takes measures to address cases of missing or inaccurate information in claims data.

4.0 Feasibility

4.1 Data Elements Generated as Byproduct of Care Processes

The data elements used in this measure are generated, collected and/or used by healthcare personnel during the provision of care (e.g., blood pressure, laboratory values, diagnosis, depression score). The data collected during care provision are then translated into the appropriate coding system (e.g., ICD-10 diagnoses, MS-DRGs) for use in Medicare claims.

4.2 Electronic Sources

All data elements are in defined fields in electronic claims.

4.3 Data Collection Strategy

4.3.1 Data Collection Strategy Difficulties

Lessons and associated modifications may be categorized into three types: data collection procedures, handling of missing data, and sampling data associated with beneficiaries who died during an episode of care.

4.3.1.1 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. If this measure is used in a CMS program, calculation and reporting would be done in line with that program’s reporting practices.

4.3.1.2 Missing Data

This measure requires complete beneficiary information, and a small number of 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 accurately adjust 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.

4.3.1.3 Sampling

During measure testing, Acumen noted that episodes in which the beneficiary died prior to the episode end date exhibited different cost distributions compared to other episodes. To avoid this effect’s potential impact on clinician scores, this measure does not include episodes for which the beneficiary’s date of death occurs prior to the end of the episode window.

5.0 Usability and Use

5.1 Use

5.1.1 Current and Planned Use

The measure was developed for potential use in the Merit-based Incentive Program (MIPS), under a contract with CMS.

5.1.2 Feedback on the Measure and Development Process

5.1.2.1 Technical Assistance Provided During Development or Implementation

Development: Field Testing

Acumen and CMS conducted a national field test of 11 episode-based cost measures developed in 2018, including the Lumbar Fusion measure, for a 35-day comment period (October 3 to November 5, 2018). We provided field test reports to a sample of clinician groups and clinicians.²⁹ Each report included information for all measures for which the clinician or clinician group was attributed 10 or more episodes. The testing sample was selected to balance coverage and reliability, since a key goal of field testing was to test the measures with as many stakeholders as possible. This sampling technique was used for field testing only and does not determine case minimums used for any potential program implementation.

- Total testing sample across 11 episode-based cost measures: 14,237 TINs; 63,984 TIN-NPIs
- Testing sample for Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels: 1,468 TINs; 3,356 TIN-NPIs

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 for each measure (comprising a Draft Cost Measure Methodology document and a Draft Measure Codes List file), a Measure Development Process document, a Frequently Asked Questions document, and a Fact Sheet.³⁰ During field testing, Acumen conducted education and outreach activities including a national webinar, office hours with specialty societies, and Help Desk support.

5.1.2.2 Technical Assistance with Results

Field Testing

During the feedback period, 2,388 field test reports for episode-based cost measures were downloaded by 403 clinician groups (TINs) and 1,985 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 field test report contained the following sheets:

- High-level summary results across all episode-based cost measures being field tested

²⁹ The field test reports were available for download from the CMS Enterprise Portal: <https://portal.cms.gov/wps/portal/unauthportal/home/>.

³⁰ 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/MACRA-Feedback.html>.

- Results for each measure including cost measure score and breakdown of episode cost compared to the national average and TIN/TIN-NPIs with a similar patient case mix (or risk profile)
- Drill-down detail for each measure, including more detailed information on potential cost drivers in the TIN/TIN-NPI's episodes. For example:
 - Analysis of utilization and cost for the measure by specific service categories (e.g., outpatient evaluation and management services, procedures, and therapy, hospital inpatient services, emergency room services, post-acute services)
 - Breakdown of costs for Physician/Supplier Part B and inpatient claims (e.g., top 5 most billed services and by risk bracket)
- Episode-level table with detailed information for all episodes attributed to the TIN/TIN-NPI across all measures in the report
 - Data across six major categories: (i) episode costs, (ii) beneficiary information, (iii) attributed clinician(s), (iv) evaluation and management visits performed during episode, (v) Physician Fee Schedule costs to Medicare billed during episode, and (vi) other providers rendering care.

A mock field test report can be viewed on the CMS MACRA Feedback webpage.³¹

Education and Outreach

Acumen directly conducted outreach via email to tens of thousands of stakeholders using the stakeholder contact list developed through previous education and outreach and clinician engagement efforts, as well as CMS, Quality Payment Program, and other available listservs. More detail on this outreach can be found in the Field Test Summary Report on the CMS MACRA Feedback webpage.

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. Acumen also hosted two office hour sessions with members of Clinical Subcommittees and workgroups to provide an update on development and field testing. Across all four office hours sessions, there were over 100 attendees.

Acumen worked with the Physician Value helpdesk and QPP Service Center to answer stakeholder questions during field testing and continued to answer questions after the feedback period ended.

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. Around 85 comments and questions were received via webinar chat and on the phone.

A post-field testing webinar was held on March 27, 2019 to provide an update on the measures following field testing. The webinar consisted of a 60 minute presentation providing an overview of the basics of measure construction, highlighting refinements made after field testing, and

³¹ CMS, "Episode-based Cost Measures Mock Field Test Report," *MACRA Feedback Page*, <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2018-Mock-report-for-Episode-Based-Cost-Measures.xlsx>.

summarizing the testing done on the measures. This presentation was followed by a Q&A session.³²

5.1.2.3 Feedback on Measure Performance and Implementation

Field Testing

In total, Acumen received 67 survey responses and 25 comment letters, including many from specialty societies representing large numbers of potentially attributed clinicians.

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

Pre-Rulemaking

CMS received 37 comments on the 11 episode-based cost measures included in the Measures Under Consideration List released in December 2018. This included four comments for the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure. After the MAP Clinician Workgroup meeting in December 2018, there was another public comment period on their preliminary recommendations, which received 23 comments across the 11 measures, with two comments specific to the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure.³³ These public comment periods were facilitated by NQF. Stakeholders were able to submit their comments via the NQF website.

5.1.2.4 Feedback from Providers being Measured

Field Testing

The Field Testing Feedback Summary Report presents all feedback gathered during the field testing period. The following list synthesizes some of the key points that were raised through the field testing feedback period:

- *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 be extended or kept open, given the large amount and complexity of the information that was presented.

³² CMS, Webinar Recordings, Slides, and Transcripts, *QPP Webinar Library*, <https://qpp.cms.gov/about/webinars>.

³³ Measure Applications Partnership, *National Quality Forum*, https://www.qualityforum.org/Setting_Priorities/Partnership/Measure_Applications_Partnership.aspx.

- *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.

The report additionally contains measure-specific feedback, which was used as the basis for the post-field testing refinements that were made to the measures, summarized below:

- Refinements to trigger codes, attribution, sub-groups, episode windows, assigned services, risk adjustment variables, exclusions, and alignment of cost with quality
- Adding/removing certain trigger codes and assigned services, further sub-grouping, and revising the attribution methodology
- Stakeholders also noted that the level of clinician engagement in the development of these episode-based cost measures is a significant improvement over the development process for earlier cost measures.

5.1.2.5 Feedback from Other Users

Pre-Rulemaking

The MAP recognized the importance of cost measures to the MIPS program and conditionally supported the Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels cost measure pending NQF endorsement. Specifically, the MAP encouraged the NQF endorsement Cost and Efficiency Standing Committee to consider the appropriateness of the risk adjustment model to ensure clinical and social risk factors are reviewed and included when appropriate. MAP cautioned about the potential stinting of care and noted that appropriate risk adjustment could help safe guard against this practice. The MAP also encouraged the Standing Committee to examine the exclusions in this measure to ensure appropriate attribution.

5.1.2.6 Consideration of Feedback

Field Testing

Careful consideration was given to all feedback gathered during field testing, and several updates were made to the measure based on the recommendations of field testing commenters and an expert clinician workgroup comprised of subject matter and measure-development experts.

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 expert clinician 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.

The changes to the Lumbar Fusion measure made after consideration of field testing analyses and stakeholder feedback are:

- **Risk Adjustment:** Add risk adjustors for:
 - Frailty variables: Osteoarthritis, Anemia, Home Oxygen, Walking Aid, Dementia, Skilled Nursing Facility Visit, Wheelchair, Home Hospital Bed

- Recent hospitalization for DRG 551: Medical Back Problems within 120 days before the trigger
- **Exclusions:** Add exclusions for:
 - Spinal Fusion Except Cervical within 120 days prior to the episode
 - Any lumbar fusions that have diagnosis codes within DRGs 456-458 (Spinal Fusion with curvature, malignancy, infections, or extensive fusions)

5.2 Usability

5.2.1 Improvement

n/a. The measures have not yet been implemented, and as such have not had influence over performance.

5.2.2 Unexpected Findings

n/a. There were no unexpected findings during the development and testing of this measure

5.2.3 Unexpected Benefits

n/a. There were no unexpected benefits during the development and testing of this measure.

6.0 Related and Competing Measures

6.1 Relation to Other Cost Measures

There are currently no related NQF-endorsed or non-NQF-endorsed cost measures that address this same measure focus or target population. There are no competing NQF-endorsed or non-endorsed cost measures that address both this same measure focus *and* at this same target population.

6.2 Harmonization

n/a

6.3 Competing Measures

n/a

Contact Information

Measure Steward Point of Contact

Organization: Centers for Medicare & Medicaid Services

Name: Joel Andress

Email Address: joel.andress@cms.hhs.gov

Phone Number: (410) 786-5237

Developer Point of Contact

Organization: Acumen, LLC

Email Address: macra-cost-measures-info@acumenllc.com

Phone Number: (650) 558-8882

Other Additional Information

Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels 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 for Degenerative Disease, 1-3 Levels 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](#).³⁴

³⁴ CMS, “Episode-Based Cost Measure Field Testing Measure Development Process,” *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>.