NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3509

De.2. Measure Title: Routine Cataract Removal with Intraocular Lens (IOL) Implantation

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

De.3. Brief Description of Measure: The Routine Cataract Removal with Intraocular Lens (IOL) Implantation cost measure evaluates clinicians' risk-adjusted cost to Medicare for beneficiaries who receive this procedure. The cost measure score is a clinician's average risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during the 60 days prior to the clinical event that opens or 'triggers' the episode, through 90 days after the trigger. Beneficiary populations eligible for the Routine Cataract Removal with IOL Implantation measure include Medicare beneficiaries enrolled in Medicare Parts A and B during the performance period.

IM.1.1. Developer Rationale: Routine cataract surgery is the most common surgical procedure in the United States, including among Medicare beneficiaries.[1] It was estimated that Medicare spends more than $3.4 billion annually on the treatment of cataracts, with cataract extraction with IOL implantation specifically as the most common procedure.[2] The Routine Cataract Removal with IOL Implantation episode-based cost measure was recommended for development by an expert clinician committee—the Ophthalmologic Disease Management Clinical Subcommittee—because of its high impact in terms of patient population and Medicare spending, and the opportunity for incentivizing cost-effective, high-quality clinical care in this area. The Clinical Subcommittee provided extensive, detailed input on this measure.

[1] Pershing, S., D. E. Morrison, and T. Hernandez-Boussard. "Cataract Surgery Complications and Revisit Rates among Three States." [In eng]. Am J Ophthalmol 171 (Nov 2016): 130-38.

[2] Brown, G. C., M. M. Brown, A. Menezes, B. G. Busbee, H. B. Lieske, and P. A. Lieke. "Cataract Surgery Cost Utility Revisited in 2012: A New Economic Paradigm." [In eng]. Ophthalmology 120, no. 12 (Dec 2013): 2367-76.

De.1. Measure Type: Cost/Resource Use

S.5. Data Source: Claims

Enrollment Data

Other

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

Criteria 1: Importance to Measure and Report

1a. High impact or high resource use:

The measure focus addresses:

− a demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality).

AND

1b. Opportunity for Improvement**:**

Demonstration of resource use or cost problems and opportunity for improvement, i.e., data demonstrating considerable variation cost or resource across providers

1a. High Impact or high resource use.

- This measure calculates the risk-adjusted cost to Medicare for beneficiaries who receive routine cataract removal with intraocular lens (IOL) implantation. The measure is specified at the individual clinician and clinician group level by calculating the clinician's average risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician.
- Medicare spends more than $3.4 billion annually on the treatment of cataracts. The intent of this measure is to capture opportunities for care improvement, specifically in mitigating costly complications as a result of the cataract surgery.

1b. Opportunity for Improvement.

- The developer provides data demonstrating that routine cataract surgery is the most common surgical procedures in the United States with a range of cost performance at the TIN and the TIN NPI level. Specifically, the interquartile range of performance for TIN level scores is $238, and mean performance of $3,041. The interquartile range of performance for TIN-NPI is $232, and mean performance of $3,038.
- The developer also provides citations demonstrating that complications from cataract surgery, specifically endophthalmitis after surgery, resulted in an 83% greater episode costs.

Questions for the Committee:

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

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

Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b)

1a. High Impact or High Resource Use

Comments:

**Yes

**Yes

**Yes – estimates that Medicare spends more than $3.4B annually on treatment of cataracts and this procedure is the most common.
**Yes
** Yes - it is high volume
** yes, measure developer describes that this is a common procedure for Medicare beneficiaries and in the aggregate represents a large amount of spending for the Medicare program. There was insufficient information (e.g., published studies) on variation in episode spending and the relationship to poor quality of care. The overall interquartile range of variation seems rather modest on a per episode basis ($232).
**Yes

1b. Opportunity for Improvement

Comments:

**Yes for both.
**Yes
**Some variation found when testing at both TIN (clinician group) and TIN/NPI (indiv. Clinician) for those with a minimum of 10 procedures in CY2017. Average cost $3k, low under $2k and high nearly $5k. Measure adopted in MIPS for CY2019 performance, so it has not yet been reported.
**Yes
** No. Even in the aggregate, the IQR is very small in terms of dollars. They do not provide the distribution of the strata, which I suspect would show even narrower opportunity for improvement. They also do not cite hypotheses or data that show which resources may be overused and at the discretion of the provider. Complication rates are known to be very low.
** In theory you could reduce the gap in spending (IQR of $232); unclear from the description is what types of quality improvements would lead to improvement on the measure. Seems a substantial contributor to spend in this area is having procedure done in hospital outpatient dept, and that is a different intervention than improving quality of care to reduce spend. Variation is rather modest (IQR of $232, and about $500 comparing the 10th to 90th percentile performance on episode cost)
**Yes, the significant variation in cost across lowest and highest cost providers and between ACS and HOPD settings strongly suggests opportunity for improvement.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing

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

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

Evaluators: Christie Teigland, Karen Joynt Maddox, Susan White, Ron Walters, Jen Perloff, Jack Needleman **(**Evaluation A: Methods Panel**)**

Methods Panel Reliability Ratings: H-3, M-3, L-0, I-0 (Moderate)
Methods Panel Validity Ratings: H-0, M-3, L-2, I-0 (Consensus Not Reached)

Measure evaluated by NQF-convened Clinical Technical Expert Panel? ☒ **Yes** ☐ **No**

Evaluators: David Vollman, Frank Burns, Kristen Carter **(**Evaluation B: Technical Expert Panel)

Reliability

2a1. Specifications:

The measure is well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability.

2a2. Reliability testing:

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

2a2. Reliability Testing:

- The developer conducted testing at the measure score-level and data element level
- Data element testing:
    - Data element testing was conducted via CMS auditing programs for Parts A &B Claims data. The developers did not provide information on confirmation of the procedure and diagnosis code.
    - The demonstration of data element validity did not meet NQF standards (i.e., description of CMS audits, fraud detection efforts)
- Measure Score
    - There were 490,714 Medicare beneficiaries included in the TIN level testing analysis and 485,216 beneficiaries included in the TIN-NPI level measure testing.
    - Measure score reliability testing included test-retest with correlations, and signal to noise.
        - Test-retest is conducted using two sets of episodes, assessing the correlation and quintile rank stability between a TIN or TIN-NPI's cost measure scores calculated from both samples; ranked clinicians by their score within each sample and stratified clinicians into quintiles; then calculated the percentage of clinicians who changed in measure score quintile between the two samples.
            - The test-retest correlation coefficients were 0.93 and 0.92 respectively for clinician groups and clinicians.
        - 70% of groups and 67% of clinicians who were in lowest spending quintile in first sample were also in lowest spending quintile in 2nd sample, and 92% of groups and 89% of clinicians were in one of two lowest.
        - The developers found a fairly high degree of consistency in the top and bottom quintiles, but much less consistency in the middle of the distribution.
    - The signal to noise analyses relied on the Adams' method (ratio of between variance to total variance). Mean reliability scores were also high, 0.95 for groups (TIN) and 0.94 clinicians (TIN-NPI).

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- Do you have any concerns with the reliability testing that was not identified by the Scientific Methods Panel?
- Would the Committee like to accept the SMP vote on reliability?

Staff Preliminary rating for reliability (based on SMP rating): ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

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

Validity

2b1. Specifications align with measure intent:

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

2b2. Validity Testing:

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

2b3. Exclusions:

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

2b4. Risk Adjustment:

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

2b5. Meaningful Differences:

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

2b6. Multiple Data Sources:

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

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

2b1. Specifications Align with Measure Intent:

- An NQF-convened Ophthalmology Technical Expert Panel (TEP) generally agreed that the clinical population was appropriate.
- One TEP member sought clarity on the use of payments to determine episode costs.
    - Developer Response: Developer clarified that standardized payments were used across the episode-based cost measures to enable fair comparisons in light of different payment rates across the country. The measure uses standardized allowed payments, or the standardized amount providers can be paid for certain services.

2b2. Validity Testing:

- **Face Validity Testing**
    - The developers used a clinical subcommittee, a technical expert panel, a person and family committee, and a national stakeholder feedback to provide input on measure and cost components attributable to this procedure episode of care measure.

- The process was structured to assure greater than 60% consensus throughout measure development, but no specific numbers presented.
- The face validity testing information provided by the developer does not meet NQF validity testing requirement requirements.
    - NQF face validity testing requires that an expert group has been convened and a systematic assessment of the measure score has been conducted. This assessment should examine whether the expert group agrees that the measure score reflects the measure intent and provides an adequate reflection of cost and resource use performance. The degree of consensus and any areas of disagreement must be provided/discussed.

- **Empirical Testing**
    - Due to the inadequacy of face validity, the SMP focused its evaluation on the empirical validity testing.
    - Empirical validity was assessed by examining correlation with other known indicators of resource utilization in administrative claims data, specifically complications related to cataract removal.
        - Correlation analysis showed expected correlation of higher cost and complications.
        - The mean observed to expected cost for episodes with services related to complications during the post-trigger period is 1.04, compared to 0.95 for episodes without services related to complications during the post-trigger period.
    - The NQF Scientific Methods Panel (SMP) reviewed the methodology for empirical validity testing and did not reach consensus.
        - Some members of the SMP expressed concern about the approach to empirical validity testing. Specifically, there was concern that the measure construct which relies on administrative claims was compared to another measure with the same data elements – which were also generated using administrative claims and used in the performance measure score. As such, the SMP members were concerned the method used by the developer did not represent correlation to an independent variable or measure.
        - Other SMP members, while acknowledging this concern, agreed that the methodology used by the developer helped to demonstrate the construct validity of the measure. They also agreed that administrative claims-based measures can be validated with other administrative claims-based measures.
    - The NQF SMP encouraged the Cost and Efficiency Standing Committee to consider the empirical testing conducted by the developer to determine if it is adequate to meet the NQF endorsement criteria.

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

- The NQF Clinical TEP generally agreed that the clinical population was appropriate, but expressed some concerns and sought clarity on the rationale for some of the specifications and decision logic.
- Exclusions eliminated a large number of patients (~40%) but left a significant number of patients left to implement measure. Developer clarified that after exclusions applied, 87.1 TIN, 83.7 TIN-PINs met case minimum of 10 after exclusions applied.
- One TEP member expressed concern over the list of included CPT codes, specifically the prophylactic treatment of retinal tears or holes. Often, even when the cataract surgery is normal, post-op tears are not related to the quality of the surgery or treatment and are unpredictable.
    - Developer Response: prophylactic treatment has not been shown to be beneficial. Risk is higher if there are complications in the surgery. If diagnosed before the surgery, they would be excluded. Not a very common occurrence.

- The developer acknowledged one of the limitation of this measure is the inability to capture Part D drug data. There are often significant costs associated with eye drops and medications associated with cataract surgery.
- One TEP members expressed concern about the inclusion of pass through drugs in the costs attributed to the physician. There was concern that these drugs may be used more frequently and the cost of these drugs is high and may impact physician use of these going forward in an effort to manage costs.
    - Developer Response: These drugs are optional and there are other treatment options for this purpose as well as available generics. The developer explained that the general approach to including costs in the measure was to capture all relevant costs and countered that physicians who use more expensive drugs should be measured as such. The developer's clinical subgroup committee co-chair noted that there has not been any published evidence that these medications are associated with lower costs or better outcomes for patients and that there are other drugs with lower costs that can be administrated at the time of surgery. Currently there is only one pass through medication included in this measure (Omidria). The subcommittee co-chair noted that pass through status will go away at some point and facilities will have to determine how to enable ongoing use of the drug once it is no longer a pass through medication. The developer noted that each pass through drug would be evaluated for inclusion in the measure individually going forward.

2b4/2c. Risk adjustment

- The Routine Cataract Removal with IOL Implantation measure is stratified into four sub-groups: ASC/Bilateral, ASC/Unilateral, HOPD/Bilateral, and HOPD/Unilateral.
    - Sub-groups for unilateral and bilateral surgery are used to account for scenarios where some services may be applied to a second surgery performed in close succession, meaning that bilateral procedures will likely be more expensive than unilateral ones.
    - The stratification for site of service accounts for access factors, as some clinicians may not have access to an ASC (which is a lower cost setting than HOPD) due to regional availability or as a result of health plan contracting arrangements.
    - These sub-groups represent more homogenous patient cohorts to enable meaningful clinical comparisons based on information available on the trigger claim. These sub-groups are useful in ensuring clinical comparability so that the corresponding cost measure fairly compares clinicians with a similar patient case-mix. A separate risk adjustment model is created for each stratified group, so that clinically meaningful distinctions in the beneficiary population are preserved.
- The developer assessed potential disparities by analyzing social risk factors of gender, dual status, income, education and unemployment.
    - The developers tested the impact of including social risk factors using T-tests and F-tests of variable coefficients and p-values, testing with step-wise regression models, and testing the final models with and without social risk factors. The developer noted that while individual bivariate testing demonstrated significance of the social factors, the inconsistent direction of the social risk factors and high correction between the measure scores with and without the social risk factors indicated that the final model sufficiently accounts for the effects of social risk factors on clinician measure scores.
    - No social factors were ultimately included in risk adjustment based on results of empirical analysis.
- Acumen's clinical subcommittee co-chair acknowledged one of the limitations of this measure in that HCCs were used as a primary risk adjuster. They were unable to correlate certain conditions in the HCC model with increased or decreased risk for ophthalmology procedures. They also questioned the applicability of this model to outpatient episodes, since it was developed for use with inpatient

episodes. HCCs are also being used for the value based modifier, but it has not been tested for validity for application for these types of measures; Ultimately, it was used in an effort to be consistent across the measurement approach for all of the episode-based measures.

- The NQF TEP Sought clarity on whether residency status (i.e., GC modifier) or the impact of resident surgery was taken into account in the risk adjustment model or site of service.
    - Developer Response: The GC modifier was included as a variable in the risk model and is applied to which ever site the resident-performed surgery took place. It was expected that resident surgeries would be longer and/or have higher risk for complications.

2b5: Meaningful Differences

- The developer assessed meaningful differences by stratifying the clinician measure scores by meaningful characteristics and investigating the clinician score distribution by percentile. Characteristics included: urban/rural, census division, census region, risk score, and the number of episodes attributed to the clinician.
    - Developer reports large performance difference among clinicians:
        - the 99th percentile of the measure score is nearly two times the 1st percentile at both the TIN level and TIN-NPI levels;
        - the Cataract measure score at the 90th percentile is approximately 20 percent greater than the score at the 10th percentile at both the TIN and TIN-NPI level; and
        - the mean Cataract score for episodes in HOPD sub-groups is approximately 30 percent higher than for episodes in ASC sub-groups at both the TIN and TIN-NPI levels.
    - There were no systematic regional difference in clinician score; clinicians in urban areas seem to perform comparably to those in rural areas.
    - Clinicians with more episodes perform similarly to those who perform fewer procedures.
    - Measure scores also show little variation by risk score decile.

Questions for the Committee regarding validity:

- **The SMP did not reach consensus based on the empirical validity testing.** Did the developer's submission adequately demonstrate empirical validity?
- Are there any concerns with the developer's approach to determining social risk factors for inclusion in the risk model?

Combined Scientific Methods Panel Preliminary Analysis of Scientific Acceptability

Measure Number: 3509

Measure Title: : Routine Cataract Removal with Intraocular Lens (IOL) Implantation

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☒ **Cost/Resource Use**

☐ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☒ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**

☒☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☐ **Registry Data**

☒☐ **Enrollment Data** ☐☒ **Other – Reviewer #1: Longterm care LDS**

Level of Analysis:

☒ **Clinician: Group/Practice** ☒ **Clinician: Individual** ☐ **Facility** ☐ **Health Plan**

☐ **Population: Community, County or City** ☐ **Population: Regional and State**

☐ **Integrated Delivery System** ☐ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

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

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

   NOTE*: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility*

   , so no need to consider these in your evaluation.

   - **Reviewer #1:** S.4 is blank – believe it should be 'Medicare Part B Beneficiaries'

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

   - **Reviewer #2:**Detailed list of cost categories and files to be accessed in separate document incorporated by reference into measure.
   - **Reviewer #3:**To achieve a homogeneous elective population, a detailed list of ten defining criteria for inclusion is provided. This was made possible by the large Medicare beneficiary population. Attribution rationale is provided for use in the MIPS QPP. The inclusion criteria have only a minimal effect on the percentage of beneficiaries of any particular demographic. The difference between beneficiaries being included or not included is minimal.
   - **Reviewer #4:**As with the other measures in this set, the assignment rules can get complex with different combinations of procedure and diagnosis codes required in different time periods. It would be helpful to see a table summarizing the different rule used for assignment. Also, the focus on the services administered by a single provider seems limited in the context of complex care processes that often involve many clinicians.

RELIABILITY: TESTING

Submission document:

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

   ☐ **Yes** ☐ **No**

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   - **Reviewer #2:**Data element: Review of CMS procedures for data auditing
     - Score level:
     - Signal to noise measurement

- Split sample correlation and analysis of consistency of quintiles across split samples

- **Reviewer #5:**Assessment of data element reliability based on references to extensive CMS auditing programs for Parts A &B Claims data. Assessment of measure score reliability was completed used two methods. The first was a test-retest approach using 2 sets of episodes and evaluation of correlation between two scores as well as change in quintle rank of the clinical group or clinician. The second approach evaluated signal to noise performance of the scores.
- **Reviewer #1:**Developer claims data element reliability testing, but does not present results. Relying on CMS claims audits to demonstrate reliability – this is weak, but not required since score level reliability is tested.
- **Reviewer #3:**At the data level, reference is made to several auditing programs in place for Medicare data and the Comprehensive Error Rate Testing Program. A reference is given. At the measure level, a test-retest approach was applied at the clinician level. Derivation of a ranked performance score in one sample was applied to the other sample to check for consistency.
- **Reviewer #6:**Test-retest and ICC; appropriate methodology.
- **Reviewer #4:**On data element reliability, the measure developers share information on CMS's financial auditing process, but provide no information on confirmation of the procedure and diagnosis code. Data element testing is incomplete.
    - On the score level testing, the measure developers use 'test-retest' and calculate a reliability score. For 'test-retest' they use a random set of episodes, but do not vary the years of data making it hard to understand if scores are consistent over time. For score reliability they use the 'Adams' ratio of between variance to total variance. This is appropriate for this measure.

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

   Submission document: Testing attachment, section 2a2.3

- **Reviewer #5:**The test/retest correlation coefficients were .93 and .92 respectively for clinician groups and clinicians. A large percentage (70% of groups and 67% of clinicians who were in lowest spending quintile in first sample were also in lowest spending quintile in 2nd sample, and 91% and 89% were in one of two lowest. This indicates strong reliability in terms of reproducing same results over different samples/timeframes. Mean reliability scores were also very high, .95 and .94 for groups and clinicians respectively.
- **Reviewer #1:**Results only stated as % more than 0.4 – apparently that is CMS' threshold? Are we required to use their threshold? It would be much easier to assess with some sort of distribution of the reliability statistics.
- **Reviewer #3:**100% of the TIN's and TIN-NPI's had a reliability score greater than 0.4, classified as "moderate". If the population is limited to those with a 10-case minimum, the mean reliability is 0.95 for TIN's and 0.94 for TIN-NPIs.
- **Reviewer #6:**Test-retest using two mutually exclusive samples, limiting to those with a 10-case minimum: Pearson correlation of 0.93 at the TIN level and 0.92 at the TIN-NPI level.
    - Ratio of between-group variance to sum of between and within-group variance (ICC): 100% of TINs and 100% of TIN-NPIs met the 0.4 cutoff; mean reliability for TINs is 0.95 and for TIN-NPIs is 0.94 using a 10-case minimum. This indicates high reliability.
- **Reviewer #4:**The reliability scores appear to be quite high. The test-rest analysis is less convincing b/c high cost outliers are likely to be randomly distributed across providers. On average the two samples look similar. However, any one provider who gets a high cost outlier may be unfairly penalized. Also, we see a fairly high degree of consistency in the top and bottom quintiles, but much less consistency in the middle of the distribution.

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

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**– **Reviewer #4:** although I would use different samples over time for the 'test-retest' analysis.

   ☐ **No**

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

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

   Submission document: Testing attachment, section 2a2.2

   ☒☐ **Yes**

   ☐☒ **No**

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

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

    ☒☐ **High** (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)

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

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

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

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

- **Reviewer #2:**Signal to noise reliability scores were high by both conventional standards and implicit standards of Adams et al re potential for misclassification.
  - The split sample test/retest results had high correlation coefficients (>.9). The proportion of cases in sample one in the lowest quintile also in the lowest quintile in sample two were 70% for TIN and 67% for TIN-NPI. The proportion of cases in sample one in the highest quintile also in the highest quintile in sample 2 were 79% and 76% respectively. Given the high correlation coefficients, this level of stability may be near the max that can be expected. **THE COMMITTEE AS A WHOLE SHOULD DISCUSS WHAT IT CONSIDERS ENOUGH STABILITY IN QUINTILES, PARTICULARLY THE HIGHEST AND LOWEST, TO CONSIDER A MEASURE RELIABLE.**
- **Reviewer #5:**Results of two methods of testing shows strong reliability of measure scores
- **Reviewer #3:**Reliabilty scores fall within the accepted norm of "moderate" by CMS standards and, as stated above, are most applicable to those with at least 10 cases. In the 10th percentile the reliability is 0.88 for TIN's and 0.869 for TIN-NPI's.
- **Reviewer #6:**High based on ICCs above.
- **Reviewer #4:**Data element testing is incomplete; measure rules are complex and hard to reproduce.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

- **Reviewer #2:** Approximately half the cateract surgeries in the initial sample are excluded from analysis with the principal reasons being PQRS Exclusion: 20/40 or Better Visual Acuity within 90 Days Following Cataract Surgery and PQRS Exclusion: Cataracts: Complications within 30 Days Following

Cataract Surgery Requiring Additional Surgical Procedures. The clinicians on the TEP and full substantive committee need to evaluate the appropriateness of these exclusions.

- **Reviewer #3:**Ten exclusions are provided and tested for their effect on O/E. Although the testing revealed similar results to the final set of episodes, for observed and risk-adjusted cost they were still excluded due to clincical considerations.
- **Reviewer #5:** None
- **Reviewer #6:**None.
- **Reviewer #4:**As the measure developers point out, many of the excluded groups look no different from the included group on mean costs or the O/E ratio – seems valuable to be inclusive and keep these cases. This would allow more providers to be included in the measure.

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

    Submission document: Testing attachment, section 2b4.

    - **Reviewer #2:**The data show a wide range of costs, and the differences are meaningful. The high R-square of the risk adjuster should compress the Observed to Expected ratios. The 90th percentile OE is about 30% higher than the 10th percentile, a substantial increase.
    - **Reviewer #5:**None. Used a stratification method and found good distribution of scores.
    - **Reviewer #3:**The risk model does demonstrate large variation in performance with a measure score at the 99th percentile being 2X than1% percentile at both the TIN and TIN-NPI level. The Cataract score at the 90th percentile was 20% greater than the score at the 10% level at both the TIN and TIN-NPI level. The Cataract score for HOPD subgroups was 30% higher than for ASC subgroups and both the TIN and TIN-NPI levels.
    - **Reviewer #6:**No concerns.
    - **Reviewer #4:**In general, there seems to be less variation in cataract surgery episode costs than we see in some other episode cost measures. I do wonder about differentiating performance in the vast middle where many providers look like the median.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    Submission document: Testing attachment, section 2b5.

    - **Reviewer #2:** NA
    - **Reviewer #3:**Not applicable
    - **Reviewer #6:**No concerns.
    - **Reviewer #4:**None.

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

    Submission document: Testing attachment, section 2b6.

    - **Reviewer #2:**Missing data are principally due to other primary payer and not continuously enrolled. These look like reasonable exclusions due to missing data but it would be good to have an explanation for the high number of not continuously enrolled beneficiaries.
    - **Reviewer #3:**The incidence of missing data is provided and consists of missing birth date, death before trigger date, primary payer other than Medicare, and non-enrollment. No concerns about the effect on validity as the measure is for cost in Medicare beneficiaries.
    - **Reviewer #6:**No concerns.
    - **Reviewer #4:**None.
    - **Reviewer #5:** None

16. **Risk Adjustment**

16a. **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐☒ **Stratification**

16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☒☐ Yes ☐ No ☒☐ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☒☐ Yes ☐☒ No ☐ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☐☒ Yes ☒☐ No

16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒☐ Yes ☐☒ No

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐☒ Yes ☒☐ No

- **Reviewer #2:**The risk adjustment analysis seems to have been done properly, with a very high R-square of 0.75. As a result the calibration, measured by comparing actual means of spending to quintiles of expected means are well calibarated. The risk adjustment model should be compressing the variation in the costs and may change relative rankings. Data on these questions is not provided.

16d.5.Appropriate risk-adjustment strategy included in the measure? ☐☒ Yes ☐ No

16e**. Assess the risk-adjustment approach**

- **Reviewer #2:**Reasonable approach, using similar methods to other CMS cost measures. General use of hcc's plus some measure specific covariates.
- **Reviewer #5:**The risk adjustment model was patterned after the CMS-HCC model. This does not necessarily provide a conceptualo rationale for the chronic conditions included in the model, but is well understood and studied.
    - For SES, they used two approaches. First, they 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 characteristic. However, the analysis also found that the directions of the effects of social risk factors are not consistent. For example, high income episodes may display higher spending for the ASC/Unilateral sub-group but lower spending for the other three sub-groups.
    - Second, they analyzed the impact of adding these social risk variables on overall model performance by looking at the differences in the ratio of observed to expected episode cost (O/E) with and without social factors in the risk adjustment model, Due to small changes in O/E ratios (+- 0.03 or less for 99% of cases), they concluded the social risk factors were captured in the HCC risk adjustments.
    - Due to the inconsistent direction and limited impact of social risk factor effects under the current risk adjustment model, they determined that the risk adjustment model sufficiently accounts for the effects of social risk factor on clinician measure scores. I did not see the rates for patients with SES vs. not and disagree that just because the model c statistic did not change that disparities may not exist. Need to see stratified results.

- **Reviewer #1:**This model includes a number of correlated variables – the impact of multi-colinearity is not addressed in the summary and could be creating anomalies in the model.
- **Reviewer #3:**The risk adjustment methodology is based on a model specifically derived for the Medicare population, routinely updated for changed in coding, and applied to specific conditions within the Medicare population. Measure-specific adjusters are selected with expert clinician input. Regression coefficients and standard errors for each of the co-variates is provided. Social risk factors were analyzed within the model and felt to be adequately captured within the model.
- **Reviewer #6:**Thorough model with HCCs as well as a number of status variables that capture ESRD, disability, and residence in long-term care, which is very helpful for picking up frailty. There are additional risk factors meant to improve face validity that come from technical expert panels, including teaching status (which I am actually not sure should be adjusted for in the setting of standardized cost analysis, since I think that component of costs is already backed out of the model).
- Social risk testing results were interesting – with the extensive HCCs, disability, and interactions, as well as prior utilization included in the model, social risk factors were largely either nonsignificant or associated with lower spending (unmet need?). There are a lot of findings in the risk adjustment table that are counterintuitive.
- **Reviewer #4:**Strong model fit (r-squared =0.75); model calibration also looks good**.**

<u>For cost/resource use measures ONLY</u>:

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

    x Yes X ☐ Somewhat ☐ No (If "Somewhat" or "No", please explain)

    - **Reviewer #4:**The included service set is rather narrow – this is likely to drive many physicians out of the measure.

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

    - **Reviewer #2:**No substantial concerns. Winsorizing to bring in outliers has been an acceptable approach. Attribution to provider billing for PCI is more straightforward than in some other cost measures.
    - **Reviewer #6:**None
    - **Reviewer #4:**I'd like to know how many cases are dropped b/c of the attribution – my guess is the number is relatively small for this episode, but this is not clear in the write up. Also, it is not clear to me why the measure splits ambulatory surgery centers from hospital outpatient departments – seems like the difference in setting could be driven by capacity in the market, patient choice or other factors. Why should these factors justify different resource use?

VALIDITY: TESTING

19. **Validity testing level: ☐☒ Measure score ☒☐ Data element ☒☐ Both**
20. **Method of establishing validity of the measure score:**

    ☐☒ Face validity

    ☒ Empirical validity testing of the measure score

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

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

    Submission document: Testing attachment, section 2b2.2

- **Reviewer #1:**They are essentially using an empirical approach to assess score variation WRT known cost drivers. I believe this is acceptable, but would like to see some statistical testing around the summary in 2b1.4
- **Reviewer #2:**TEP and multiple stakeholder panels, such as person and family committee, for face validity
    - Correlation of results with costs associated with complications
- **Reviewer #5**:Used multiple TEPS to assess face validity and provide input on measure and cost components attributable to the procedure episode of care. Empirical validity was assessed by examining correlation with real indicators of resource utilization, specifically complications related to cataract removal.
- **Reviewer #3:**Correlation testing with known indicators of resource or service ulilization, specifically complications related to cataract removal was performed. The mean observed to expected cost was 1.00. The mean observed to expected cost for services related to complications was 1.04, compared to 0.95 for episodes without services related to complications.
- **Reviewer #6:**R-squared. Predictive ratios by decile (discrimination)
- **Reviewer #4:**The work with clinical experts to define the measure was comprehensive. The empirical validation of the resource use measure was rather limited, comparing the score for those with and without complications. It would be more convincing to use a different measure of resource use, such as a delivery system's internal cost accounting.

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

    Submission document: Testing attachment, section 2b2.3

    - **Reviewer #2:**TEP, stakeholder panels, and face validity. Process described as being structured to assure greater than 60% consensus, but no specific consensus numbers presented. No discussion of whether panels received and how they reacted to reliability and test/retest analyses.
        - Correlation analysis showed expected correlation of higher cost and complications.
    - **Reviewer #5:**Do not see results of TEP face validity testing? The mean observed to expected cost for all episodes is 1.00, for episodes with services related to complications 1.04 and for those related to episodes without complications .95. These do not represent very large differences in costs with and without complications.
    - **Reviewer #1:**There are no statistical results to assess.
    - **Reviewer #3:**The overall R-squared for the cost measure was 0.75 with an adjusted value of 0.75. Calibration demonstrated that the average observed to predicted observed was generally close to 1.
    - **Reviewer #6:**R-squared was 0.75, adjusted R-squared 0.75. Predictive ratios by decile (discrimination): Appendix table shows good discrimination
    - **Reviewer #4:**The face validity is good, but the empirical validity appears rather weak given that it is only one comparison on a measure that specifically includes some complication costs.

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

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

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

☒☐ **Yes**

☒ **No Reviewer #2:**Reliance on CMS testing and auditing methods is reasonable but doesn't qualify as testing data elements.☒☐ **No**

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

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

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

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

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

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

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

- **Reviewer #2:**The measure is reasonably well conceived and constructed measure. Substantive experts should review the exclusion criteria, which wind up excluding half the cases.
- **Reviewer #5:**Validity results show high predictive power and good range of scores to differentiate performance on this measure. Model had high R-Square of .75.
- **Reviewer #3:**Rationale is clear, testing is appropriate, and the results are very good for model development and testing
- **Reviewer #6:**Well-explained, well-calibrated measure. Concern with whether or not quality of care provided has anything to do with the outcome (since not all costs are bad), but that's an issue for the content committee. Have a concern about adjusting for teaching status.
- **Reviewer #4:**Although face validity is important, further empirical validity testing is in order.

ADDITIONAL RECOMMENDATIONS

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

- **Reviewer #2:**As noted above, the measure is a reasonably well conceived and constructed measure, and therefore the validity appears sufficient.
  - The committee's substantive experts should assess the specific exclusion decisions, as these cut the data in half while the empirical analysis suggests similar distributions of costs of the excluded cases.
  - Reliability appears sufficient. Numbers on S-to-N are high as is rest/retest correlation coefficient. Based on standards in the literature, these are numbers that are considered indicators of high reliability. But these measures are often used in ranking with sharp cutoffs for penalties and rewards. And about 20% of sample 1 highest or lowest quintiles are not highest or lowest in sample 2. I would like the full methods committee to discuss what level of quintile instability is acceptable to judge a measure reliable.

Evaluation B: Technical Expert Panel (Preliminary Evaluation Comments)

Measure Number: 3509

Measure Title: Routine Cataract Removal with Intraocular Lens (IOL) Implantation

Type of Measure: Cost/Resource Use

---

1. Clinical Logic Evaluation of Measure (questions S8.1.-S.8.6 in submission form)

1a. To what extent is the measure population clinically appropriate?

- **TEP Member 1:** The measure population is clinically appropriate because it limits the cases to routine cataract surgery in a Medicare only population by including appropriate exclusions of clinical diagnosis and additional payors.
- **TEP Member 2:** The measure population is appropriate. The Medicare population is the right age range and a high utilizer of cataract surgery.

1b. To what extent are the definitions used to identify the measure population clinically consistent with the intent of the measure?

- **TEP Member 1:** They are consistent because the goal is to provide clinicians with resource utilization information for routine cataract surgery so appropriate exclusions (S.9.1) are used to limit the population from both the ocular co-morbidities perspective as well as limiting cost attribution errors.
- **TEP Member 2:** The definitions are clinically consistent since the episode has to be completed within a calendar year to enable an accurate lookback period and completion of the episode.

1c. To what extent does the submission adequately describe the evidence that supports the decisions/logic for grouping claims (i.e., identifying the measure population, exclusions) to measure the clinical condition for the episode?

- **TEP Member 1:** Most of the evidence relies on the clinical expert opinion of the group assembled by Accumen. There is some evidence provided about the impact of cost awareness on physician decision making but this is not as directly impactful on defining the measure.
- **TEP Member 2:** The submission adequately identifies the population, and some exclusions. It does not list the ocular comorbidity exclusions anywhere in the document. It mentions that there are some, but does not list them. The ocular comorbidity exclusions are extremely important and are not listed. For example, a patient with exudative wet macular degeneration will continue to receive anti-VEGF injections in the postoperative period which may be extremely expensive and entirely unrelated to the cataract surgery.

1d. Given the condition being measured, and the intent of the measure, describe the alignment of the length of the episode (including what triggers the start and end) with the clinical course of this condition.

- **TEP Member 1:** The episode length does make sense with the 90-day post-trigger date aligning with the traditional 90 days post-surgical global period in Medicare. The 60 days pre-trigger date look back is consistent with the timeframe that the majority of pre-operative testing and office visits directly associated with the condition would occur.
- **TEP Member 2:** The trigger date, as well as the start date and end date are appropriate in length. Most patients are done with their postoperative care within three months and see the doctor within 60 days prior to cataract surgery. The lookback period is very important to consider other diagnoses that may exclude the patient. The length of this period is also appropriate.

2. Adjustments for Comparability-Inclusion/Exclusion Criteria (question S.9.1. in submission form)

2a. Describe the clinical relevancy of the exclusions to narrowing the target population for the episode, condition/clinical course or co-occurring conditions, and measure intent.

- **TEP Member 1:** The exclusions are consistent if we are trying to limit the measure intent to only patients experiencing routine cataract surgery in a Medicare population. The exclusions accomplish this by eliminating patients that have secondary insurance coverage. From a clinical care perspective, the elimination of co-morbid ocular conditions that could impact resource use (e.g. diabetic retinopathy or glaucoma) makes sense since these patients can have more unpredictable post-operative courses that could require additional interventions and more resource use. Also, the site of service limitation to ASCs and HOPDs and comparing apples to apples populations makes sense.
- **TEP Member 2:** There are no ocular comorbidity exclusions listed, which are extremely important as discussed in my answer to 1c.

2b. Do the exclusions represent a large number or proportion of patients?

- **TEP Member 1:** Given the volumes of cataract surgery performed in the Medicare population, these exclusions should not narrow the population of patients to a significant degree as to still be able to ascertain a difference in resource utilization among providers.
- **TEP Member 2:** The exclusions do not represent a large proportion of the patients. I cannot comment on the ocular comorbidities as these are not listed.

2c. To what extent are the relevant conditions represented in the codes listed in the submission for clinical inclusions and exclusions?

- **TEP Member 1:** Unless I am missing something, I do not see any specific codes listed.
- **TEP Member 2:** The clinical inclusions and exclusions that I have seen are appropriate. Again, I have not seen the ocular comorbidity list, so I cannot comment on these. Please see my answer to 1c.

3. Adjustments for Comparability-Risk Adjustment (question S.9.3. in submission form)

3a. To what extent are the covariates (factors) included in the risk-adjustment model clinically relevant and consistent with the measure's intent? Are there other clinical factors or comorbidities that should be considered for inclusion in the model? Excluded from the model?

- **TEP Member 1:** The four risk stratifications by location (ASC/HOPD) and number of eyes (unilateral/bilateral) makes clinical sense based on the fact that HOPDs will typically be more resource intensive than ASCs and bilateral surgery will inherently cost more than unilateral surgery. The section mentions a separate risk adjustment model for each of the four stratifications but does not explicitly list variables included in those models so it is hard to comment on the what other factors should be included or excluded.
- **TEP Member 2:** The factors included are relevant and important.
- Another important factor that is not mentioned is prescription eye drops. Prescription eye drops are prescribed at the time of cataract surgery and can be extremely expensive. The cost typically runs about $80 for two generic drops and up to over $400 for brand name drops. This is hard to include as Part D is highly variable, and doctors do not control the insurance formulary. However, it can be a significant expense at the time of cataract surgery.
- Also, please see my answer to 1c.

Committee Pre-evaluation Comments:
Scientific Acceptability of Measure Properties including 2a and 2b

2a1. Reliability – Specifications

Comments:

**I have no major concerns.
**Would like physician expert opinion on the exclusions for ocular comorbidities
**n/a
**Well done, high reliability measure
** none. All episode measures are difficult to replicate, but that is not a specific issue here.
** One of the reviewers mentioned they didn't see ocular comorbidity exlusions specified and that this would be important. Overall given this is claims based measure, it could be consistently implemented. Unclear why Part D claims for eye medications is omitted.
**No concerns

2a2. Reliability – Testing

Comments:

**I have no major concerns.
**n/a
**No concerns
** there was no result data provided at the strata level, which is the level of risk adjustment and reporting.
** concern about statement that says 100% of TINs had reliability score of 0.40 or higher which was moderate. 0.40-0.60 is considered very low relability, and 0.6-.7 is low. overall mean reliability was at .95 which is good, but not surprising given large number of cases per TIN and likely small within TIN variation. Did test-retest in single year's data (which would be stronger test of reliability to assess consistency in scores across periods)
**The difference in reliability between a 10 case, 20 case, and 30 case minimum is greatest for the 10th%ile. I would like to hear more as to how this impacts the measure's reliability.

2b1. Specifications align with measure intent

Comments:

****I have no major concerns.
****Measure only for Medicare FFS beneficiaries – possible concern that roughly 30% of Medicare beneficiaries (those covered by Part C MA plans) are not included and thus do not have a complete picture of resource use for the population.
**Well described. Excellent discrimination and calibration.
** no issues
** the testing for SES (as described) was unclear regarding whether they were adjusting for between or within variation by SES factors. It isn't surprising that the various factors work in different directions, and this raised concern for me about lack of conceptual model and consideration of why different factors might result in differential effects. I don't feel the assessment of SES factors is fully complete.
**In Section 2b1.3, the correlation between the themes of Complications and Office-based Procedures and episode cost seems moderate, not strong. Whether this cost measure penalizes clinicians with higher complication rates will depend on how the information is used. In Section 2b2.2, Table 1, the impact of ocular co-morbidities is higher at the lowest and highest deciles. I would like to hear the measure developers thoughts on what caused this pattern.

2b2. Validity –Testing

Comments:

2b3. Exclusions

Comments:

**I have no major concerns.
**Want clinician input to exclusions for ocular comorbidities.
**none

**Well justified.
** no issues
**none
**Exclusions were explained adequately.

2b4/2c. Risk Adjustment/Stratification/Disparities

Comments:

**I have no major concerns.
**The inconsistency of the direction of social risk factors in risk adjustment is the principal basis for excluding from risk adjustment. I concur. Rest of risk adjustment model is standard CMS approach.
**Long-term care and disability status appear to be the only social risk factors included in the risk adjustment model; DOB is listed, would want confirmation that age is use for risk adjustment. One big risk factor for cataracts is sun exposure so I wonder if a geographic adjustment is sensible or feasible? No – it seems the rationale is implicitly because the measure relies upon claims data, which lacks robust data elements on social risk factors
**Social risk factors are appropriately dealt with.
** no issues
** I did not see a conceptual model for the adjusters they tested (per SES). Clinical adjusters seemed appropriate.
**The risk stratification methodology is explained in detail, but I would like to hear from committee members with more more expertise in this area consider the method used to be valid.

2b5. Meaningful Differences

Comments:

**I have no major concerns.
**Okay.
**none
** this is my greatest area of concern. If the only data you can show is the difference between the 1% and the 99% to show meaningful differences, then it is worthwhile questioning the value of the measure. The difference between the 10th decile and the 90th decile is also below $500 even without stratification. With stratification it could be extrememly low but is not shown
** The results from their national field test suggest providers had mixed opinions about the results and whether findings were meaningful.
**No concerns

2b7. Missing Data

Comments:

****I have no major concerns.
**none
**The developer assume that Medicare data is complete. This may not be valid. Exclusions are well justified.
**** no issues
**** This doesn't seem a problem (save for consideration of whether Part D data should be included). Rationale for excluding Part D data--I did not see it.
**No concerns.

Other Threats to Validity: Carve-outs/Attribution/Truncation/Costing Approach

Comments:

**I have no major concerns.
**None.
**It seems this measure's approach is aspirational and would need to see actual data to determine whether it appropriately measures costs that the accountable clinician/group has reasonable control of. There are 4 sub-groups (ACS bilateral and unilateral, and HOPD bilateral and unilateral) and outliers are excluded separately for each group – this seems like a wise approach to ensure outliers for one sub-group do not influence the outliers for another.

**No concerns. Generally well validated measure.
** no issues
** A key question is how much physicians are controlling differences in observed spending. Again this relates to the underlying literature showing a relationship between physician quality performance and spending for cataract surgery (in the submission package). This literature may exist, but the importance section fell short in describing this. Use of standardized prices is correct approach. The underlying assumption of this measure is that physician quality (say via complications) is leading to higher costs within an episode. The evidence in the NQF submission package was limited in that regard.
**No concerns.

Criterion 3. Feasibility

3. Feasibility

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

- All data elements are in defined fields in electronic claims
- There are no fees, licensing, or requirements to use the measure.
- Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
- Generated by and used by healthcare personnel during the provision of care, e.g. blood pressure, lab value, medical condition
- This measure uses variables from claims data submitted by healthcare providers to a Medicare Administrative Contractor and are subsequently added to the Common Working File maintained at the CMS Baltimore Data Center.
- 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.

Questions for the Committee:

- Are there any concerns regarding feasibility?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:
**I have no major concerns.
**Claims based measure. No concerns about feasibility
**none
**Certainly is feasible.
** no issues
** no issues--drives of electronic claims
**No concerns.

Criterion 4: Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

Use

4a.1. Accountability and Transparency.

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

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

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

4a1. Current uses of the measure

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

  OR
- **Planned use in an accountability program?** ☐ **Yes** ☐ **No**

Accountability program details

- Quality Payment Program - Merit-based Incentive Payment System

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

- Acumen and CMS facilitated feedback on the measure through multiple channels including a national field test of episode-based cost measures; field test reports were provided to 4,434 TINs and 7,690 TIN-NPIs. They also hosted office hours to answer questions from potential measure users, hosted National Provider Calls to engage clinicians, collected comments via several public commenting periods, and deployed online surveys to solicit feedback.
- The measure entities (clincians and clinician groups) and other stakehodlers or interested partieis submit questions or comments about the measure through an email inbox (be macra-episode-based-cost-measures-info@acumenllc.com) and could review a mock field test report posted on the CMS website.
- While some stakeholders believed the field test report presented useful information for understanding clinician cost measure performance, they also highlighted areas for improvement in regard to providing actionable information.
- Stakeholder feedback received on the supplemental field testing materials was mixed, with some stakeholders finding them helpful and informative and others believing the materials were too complex.

Additional Feedback received through the Measure Application Parternship (MAP)Process

MAP Recommendation: Conditional Support

Public and Member comments:

- While the American Medical Association (AMA) is supportive of the collaborative process CMS has used in the development of these measures, we do not believe that they were ready for MAP consideration and did not receive adequate vetting by the Clinician Workgroup. There is also a consistent problem with the timeline to provide comments back to the MAP which greatly jeopardizes the integrity of the MAP process. The AMA is troubled over the lack of transparency and inconsistent application of the Measure Selection Criteria (MSC) to these episode-based cost measures. Specifically, no information regarding the individual measure specifications, attribution methodology, or reliability and validity testing results were released for member and public review prior to the MAP Clinician Workgroup meeting, modifications to the measures based on preliminary feedback are still being made, and, to our knowledge, the Workgroup members did not have any detailed information in front of them at the time of the discussion. The developer only cited some limited information on how the measures were developed and tested. Given the degree of interest from numerous medical specialty societies, the AMA looked forward to a robust and detailed discussion on each of these cost measures but unfortunately, it did not occur.
- ASCRS thanks the MAP for its comments in support of appropriate risk adjustment and the need for an alternative metric than a national average to compare a physician's cost. We believe the exclusions submitted in the cataract episode methodology are key factors in assuring physicians are not held accountable for the cost of treating patients with co-morbidities, which is out of their control. In addition, we believe an appropriate alternative to a national average cost would be to compare the physician's expected cost versus the physician's observed cost. Given the wide range in average costs for each of the episode sub-groups noted during the recent field test of this measure, the national average cost of cataract surgery identified, $2,676, is a misleading number to ophthalmologists. We suggest the measure be based on comparing physicians to how far they diverge from the expected cost, based on their mix of episode sub-groups and risk percentile.
- Overall, we support the MAP's recommendation. We urge the measure developers to address fees that are out of the physician's control. Physicians who perform their procedures at free standing surgical centers would incur regulated fees that differ from the regulated fees in hospitals. The variation in these fees is not affected by the physician's performance.

***Questions for the Committee**:*

- Given the concerns raised in the feedback on this measure to date, does the Committee have any concerns with its use?

Staff preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

Usability

4b.1 Improvement.

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high quality, efficient healthcare for individuals or populations.

4b2. Benefits vs. harms.

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

4b3. Transparency.

Data and result detail are maintained such that the resource use measure, including the clinical and construction logic for a defined unit of measurement, can be deconstructed to facilitate transparency and understanding.

4b1. Improvement results

- This is a new measure. The developer did not provide any data to demonstrate any improvement.

4b2. Unintended consequences

- The developer did not identify any unintended consequences during measure development and testing.

4b2.Potential harms

- The developer did not identify any potential harms.

4b3. Transparency

- Stakeholder feedback received on the supplemental field testing materials was mixed, with some stakeholders finding them helpful and informative and others believing the materials were too complex.

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- What benefits, potential harms or unintended consequences should be considered?
- Do the benefits of the measure outweigh any potential unintended consequences?
- Do the measure specifications and accompanying documentation enable adequate transparency to facilitate understanding of how the measure results are generated?

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

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

Comments:

**Yes.
**Didn't see it in testing document.
**Not yet, but assume will be (as part of MIPS public reporting); Beginning to be used but not yet publicly reported; not yet disclosed.
**Not was well documented as it might be. It will be straight forward to use.
**no issues
**CMS is using in the MIPS program starting in 2019
**Via the MIPS program. No concerns.

4a2. Use – Feedback

Comments:

**I have no major concerns.
**I believe there was a TEP for the measure, but this needs discussion.
**none
**No feedback noted.
**There have been extensive opportunities for feedback.
**the picture here was mixed based on comment the measure developer received. there were recommendations provided to facilitate use of the information.

**No concerns

4b1. Usability – Improvement

Comments:

**Yes.

**Not adequately described.

**? Not really, seems to assume that measurement will automatically drive improvement (through the information alone?), and that it cost could capture consequences of care (such as cost for complications); relies on general premise that the most common procedure being measured for cost will lead to higher

**Usability not well defined. Should be straight forward.

**The measure fails for me on this dimension

**no 2nd year or 3rd year data available to gauge--this is a new measure

**Incorporation into MIPS for the 2019 measurement year suggests the possibility of improving cost, the full impact will not be known until data from subsequent years is obtained.

4b2. Usability – Benefits vs. harms

Comments:

**I have no major concerns.

**One concern noted in testing document is that additional cost of patient visits and consultation will discourage appropriate prework. Developers minimize this concern. We should look at the relative magnitudes and discuss how strong the incentive is.

**Care stinting, especially due to the lack of broader adjustment for social risk factors

**None particular to this measure. There is always danger.

**the benefits are low so the main harm is the cost of producing the measure

**no potential harms noted.

**No concerns

4b3. Transparency

Comments:

**Yes.

**yes.

**none

**Yes.

**The lack of results data by strata is concerning

**there is a lot of technical detail provided to explain how measure was constructed and how to interpret (in physician feedback/transparency reports). Unclear whether/how consumers would use.

**Yes

Criterion 5: Related and Competing Measures

- There are no related or competing measures

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: June 4, 2019

The American Society of Cataract and Refractive Surgery (ASCRS) is a medical specialty society representing nearly 9,000 ophthalmologists in the United States and abroad who share a particular interest in cataract and refractive surgical care.

Thank you for the opportunity to provide comments on the pending endorsement of the cost measure, Routine Cataract Removal with Intraocular Lens (IOL) Implantation (cataract episode measure). ASCRS opposes NQF endorsement of this measure unless it excludes any FDA-approved Medicare Part B drug on pass-through

status administered during cataract surgery. Specifically, we are concerned that if pass-through drugs are included in the cataract episode measure, it will disincentivize surgeons from using the drugs and negatively impact the utilization data CMS is collecting for this purpose. Ultimately, including Medicare Part B drugs on pass-through in the episode measure defeats the purpose of pass-through.

In addition to our concerns related to the inclusion of pass-through drugs in this measure, we are providing comments on the following elements:

- **Support for the trigger code, exclusions for risk adjustment, and the sub-groups of the episode measure.** These measure specifications were developed with the input of ASCRS and other ophthalmologists through Acumen's technical expert panel (TEP), of which I was a member. These factors reflect the TEP's goal of only measuring physicians on the costs they can influence.
- **Ongoing concern that when physicians are scored on this measure they are given an adjusted cost, expressed as a dollar figure that may differ considerably from the actual cost, and then compared to total national average cost. We recommend the cost score be expressed as the difference between the physician's expected cost versus the physician's observed cost.**

EXCLUDE DRUGS ON PASS-THROUGH

- **ASCRS urges NQF to withhold its endorsement of the cataract surgery episode-based cost measure unless CMS and Acumen remove the current FDA-approved drug administered during cataract surgery included in the measure on pass-through, and signal that any drug that has since come onto the market and is paid on pass-through, or will come onto the market, will not be included in the measure.**
- **Pass-through status is a vital tool in ensuring that new and innovative drugs are introduced to the market and is used by CMS in the formula to calculate the increase in the ambulatory payment classification (APC) group, which is the facility fee for the procedure and other related procedures, to account for the drug.** Pass-through status helps introduce a new drug into the marketplace that is used during or immediately after surgical procedures with an average estimated cost that exceeds a certain percentage of the procedure's ambulatory payment classification (APC) payment amount. It is initially put on pass-through status and paid separately for up to three years under Medicare Part B. This encourages the use of new drugs in the facility by allowing physicians time to become familiar with their use without their adding to facility cost. Separate payment for pass-through drugs is also essential to ASCs, in particular, because their lower facility reimbursements would make it difficult to afford new, high-cost drugs. Over this time period, CMS measures the utilization of the drug and, when the drug goes off pass-through status, adjusts the reimbursement level for the bundled facility fee based on the utilization data gathered and the formula. **Pass-through status is intended as a means for CMS to gather data not influenced by other factors. If drugs on pass-through status are included in the measure, physicians mindful of their score on the cataract surgery measure may modify their use of the drug for reasons other than clinical appropriateness, and thus impact the gathering of utilization data, thereby defeating the purpose of pass-through.**
- **Currently, there are several ophthalmic drugs that have either recently been approved or will be approved in the near future for use during cataract surgery. One such drug—injection, phenylephrine and ketorolac, 4 ml vial—is included in the measure for 2019.** Specifically, these new FDA-approved drugs administered during cataract surgery that are on now on pass-through, or soon will be, have a post-operative indication, such as post-operative pain and inflammation and/or other sequela of the surgery, and eliminate the need for some or all post-operative eye drops. Reducing or eliminating the need for post-operative eye drops, which are currently furnished under Medicare Part D, represents a substantial cost-saving both to the Medicare program and the patient. In addition, eliminating the need for post-operative eye drops improves patient compliance and leads to better clinical outcomes. However, since Part D costs are not a factor in the cataract episode measure, using these Medicare Part B pass-through medications during cataract surgery and including them in the episode calculation would increase the total episode cost and would inaccurately designate the surgeon as high-cost. **Beyond the primary goal of preserving pass-through status to ensure accurate utilization calculations, we believe including these drugs with a post-operative indication on pass-through would go against the goal of the episode-based cost measures of encouraging physicians to make more efficient use of resources.**

ASCRS believes that episode-based cost measures are a more effective method of measuring clinician resource use than population-based measures because they only include the costs of care that are within the physician's control. However, physicians have no control over the cost of drugs as they enter the market, and therefore, including the cost of these drugs in the measure is contrary to the goals of episodic-based measurement. To ensure that clinicians are not penalized for using drugs on pass-through and that pass-through status is preserved to collect accurate, market-based utilization data, we recommend that any FDA-approved Medicare Part B drug administered during, or at the end of, cataract surgery that is on pass-through status be excluded from the cataract surgery episode-based cost measure. NQF should withhold its endorsement of the measure until there are assurances from CMS that the current drug on pass-through included in the measure, and any other pass-through drug that has entered the market or will in the future, be excluded.

TRIGGER CODE

- **ASCRS supports the use of CPT code 66984 as a trigger for the cataract episode measure.** Routine cataract removal with 66984 requires homogeneous and comparable resources for nearly all patients. As a high-volume code, it will provide enough data to identify outlier physicians who are practicing outside of established patterns. **66984 is the only code included in the specifications as a trigger code. No other codes should be considered.**
- ASCRS does not support including other codes, such as complex cataract surgery, 66982, in this measure, as it will not yield comparable enough data to measure a physician's resource use accurately. Patients undergoing cataract surgery that requires the use of the complex cataract code may suffer from a wide variety of ocular co-morbidities, or other non-ocular co-morbidities, which could require varying levels of resource use depending on the condition. For example, patients taking Tamsulosin or similar medications very frequently require the use of iris retractors, leading to the use of code 66982 instead of the usual 66984. Furthermore, these patients often have complications requiring further surgery, such as a vitrectomy. **Complex cataract surgery may require additional supplies and increases the likelihood of potential complications, resulting in a range in value too significant to provide a homogenous patient group for cost measures and, therefore, should not be used as a trigger code.**

EXCLUSIONS

- **ASCRS supports the use of criteria to exclude patients with significant ocular co-morbidities from the cost measure.** We support the use of the exclusionary criteria from quality measure 191, Cataracts: 20/40 or Better Visual Acuity within 90 Days Following Cataract Surgery. The criteria exclude patients with documentation of significant ocular conditions. The exclusions include such chronic conditions as diabetic retinopathy, macular degeneration, and glaucoma.
- **The rationale for excluding these patients is that since patients with significant ocular co-morbidities are excluded from quality measurement, they should also be excluded from cost measurement.** The developers of quality measure 191 excluded these patients because ocular co-morbidities play a significant role in whether the patient will have a predictably good
- outcome, and whether complications may arise. Surgeons do not have control over a patient's co-morbidities and should not be held accountable for additional costs in an episode if a patient suffers from one of these conditions. **If physicians are not measured on the quality outcomes of these surgeries, therefore, they should not be held responsible for the cost of these surgeries.**
- **These exclusions ensure a greater level of risk adjustment than has previously been incorporated in cost measures, such as the current measures total cost per capita and Medicare spending per beneficiary.** While we are aware that CMS includes a basic level of risk adjustment calculation for existing cost measures, and for the episode-based measures, CMS has not been able to demonstrate that its risk adjustment properly reflects the cost of cataract surgery. This risk adjustment includes factors such as beneficiary age, dual-eligibility, and some co-morbidities, but does not include ocular co-morbidities. Cataract surgery is performed on a relatively older patient base, and while some systemic co-morbidities may require additional resource use as discussed above, ocular co-morbidities play a much larger role in determining the resource use and likelihood of a good outcome than do the factors used in CMS' current risk adjustment. **Using the exclusionary criteria from the cataract quality measure is a much more**

accurate means of risk adjustment to ensure that physicians are not held accountable for the cost of care related to factors outside their control.

EPISODE SUB-GROUPS

- **ASCRS supports the inclusion of the measure sub-groups, which relate to site of service and laterality.** We believe the four sub-groups in the cataract episode measure represent significant differences in the cost of cataract surgery—largely related to Medicare's own billing and reimbursement policies—and should be separated for basis of comparison in this measure.
- **Cataract surgery can be performed in either hospital outpatient departments (HOPDs) or ambulatory surgery centers (ASCs), with reimbursements for ASCs well below HOPDs.** Cataract surgery is reimbursed 45% less in the ASC than in the HOPD. While some ophthalmologists have the option of building and owning their own ASC, some states with certificate of public need laws limit the number of existing ASCs or prevent physicians from opening new ASCs, so they may be forced to operate in HOPDs. In addition, some physicians, especially solo practitioners, may not have the resources to construct and manage their own ASC, and must operate in whatever facility, either ASC or HOPD, is available. Despite these limitations and given the choice, ophthalmic surgeons would likely prefer to operate in the lower-cost ASC. ASCs are not subject to the same requirements as HOPDs, such as extensive pre-operative testing, that are not relevant to treating ophthalmic disease. In addition, patients may prefer to undergo surgery in ASCs, which are generally easier to navigate since they are smaller, less intimidating, and have shorter wait times. Ophthalmic surgeons want to make the cost-effective choice but cannot always do so. **Given that, the episodes must include sub-groups for ASCs and HOPDs, since the site of service is not always within the physician's control.**
- **ASCRS supports the sub-groups for laterality because they reflect whether the surgeon removed cataracts in either one or both eyes during the episode window.** Patients frequently
- Develop cataracts in both eyes, and while both eyes are rarely operated on at the same time, many patients find it convenient to have the second surgery shortly after the first, usually still within the 90-day global post-operative period. Medicare has specific billing rules for physicians performing multiple procedures on the same patient related to pre- and post-operative care; therefore, the expected cost of cataract surgery performed during the global period of previous cataract surgery would be substantially different from two surgeries performed more than 90 days apart.

MEASURE SCORE BASED ON NATIONAL AVERAGE PRICE

As noted above, we support the measure sub-groups, which were determined by the clinical TEP. The physicians on the TEP selected those sub-groups to separate and avoid comparing surgeries where fundamentally different factors are contributing to the amount Medicare is reimbursing in total for the episode that are not always in the control of the physician. Given that fact, it is confusing to the physician receiving feedback on cost performance to see a dollar figure representing his or her "average" cost compared to a national average. **We believe physicians would better understand their performance on the measure and be able to take action in response if they were shown how their average observed cost compared to their average expected cost.**

- **The cataract episode should compare each physician or TIN's average observed cost to that same physician or TIN's average expected cost and not compare overall to a national average.** The clinical TEP determined that the sub-groups broadly represent the main drivers of cost in relation to cataract surgery. CMS and Acumen will determine risk-adjusted expected costs for each of the sub-groups. Each of a physician's attributed episodes' observed costs will then be compared to the expected cost for the respective sub-group and assigned a ratio to represent the divergence between the expected and observed costs. The ratios for each of the episodes are then averaged to determine the frequency of the physician's divergence from the expected cost. **We support this approach, as it ensures that the varying costs of the sub-groups outside of the physician's control, such as the facility fee, are not impacting the physician's score.**
- **However, the steps following the calculation of the average ratio should be re-thought to make the final average cost more meaningful to the physician or group practice.** Following the above calculation, the ratios of observed and expected costs are converted back to a dollar figure and compared to a national

average. When draft field test reports of the measure were distributed in late 2017, we heard from several of our members who reviewed their feedback reports and questioned how their final average cost and the national average costs were determined, and how it relates to the actual reimbursement they received for the surgery. In addition, physicians are aware that geographic differences contribute to the reimbursement level and may question why they should be compared to a national dollar average. While we understand that supplemental methodology documents are available and discuss this process, CMS does not generally distribute these explanatory resources in conjunction with performance feedback, and if a clinician does find them posted on a separate web page, they may be difficult to understand.

- **To overcome this issue, we recommend the physician's score be based on his or her own expected and observed costs, and not based on a misleading national average.** For example, if a surgeon's case mix means that 70% of his or her surgeries are unilateral and performed in the ASC, and 30% are unilateral and performed in the HOPD, then his or her expected cost should be a weighted average of those two sub-groups. Then, the surgeon's actual observed costs are compared to that expected average of a 70/30 mix of those two sub-groups. The average cost for this surgeon would be substantially different from a physician who performed 90% of surgeries in an HOPD and 10% in an ASC. **The surgeon would then be evaluated on the extent, above or below, that he or she deviates from the expected cost for his or her specific case mix. We believe this is a much more useful value to a physician than a national average.**

CONCLUSION

While ASCRS believes that episode-based measures are a more accurate means of measuring resource use because they only include the costs that are within the clinician's control, we oppose NQF endorsement of this measure until CMS and Acumen remove the current drug on pass-through from the measure and indicate that no other pass-through drug will be included in the future. We are deeply concerned that including these drugs will improperly influence utilization data collection and create an adverse incentive for physicians to avoid using innovative drugs that improve patient outcomes because they will negatively impact cost scores. If CMS and Acumen make these changes, the measure should move forward for endorsement with the current specifications for trigger code, sub-groups, and exclusions. We continue to recommend that the final score of the measure not be compared to one national average figure.

Importance to Measure and Report

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

IM.1. Opportunity for Improvement

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

Routine cataract surgery is the most common surgical procedure in the United States, including among Medicare beneficiaries.[1] It was estimated that Medicare spends more than $3.4 billion annually on the treatment of cataracts, with cataract extraction with IOL implantation specifically as the most common procedure.[2] The Routine Cataract Removal with IOL Implantation episode-based cost measure was recommended for development by an expert clinician committee—the Ophthalmologic Disease Management Clinical Subcommittee—because of its high impact in terms of patient population and Medicare spending, and the opportunity for incentivizing cost-effective, high-quality clinical care in this area. The Clinical Subcommittee provided extensive, detailed input on this measure.

[1] Pershing, S., D. E. Morrison, and T. Hernandez-Boussard. "Cataract Surgery Complications and Revisit Rates among Three States." [In eng]. Am J Ophthalmol 171 (Nov 2016): 130-38.

[2] Brown, G. C., M. M. Brown, A. Menezes, B. G. Busbee, H. B. Lieske, and P. A. Lieke. "Cataract Surgery Cost Utility Revisited in 2012: A New Economic Paradigm." [In eng]. Ophthalmology 120, no. 12 (Dec 2013): 2367-76.

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

Performance scores are provided for 4,515 clinician group practices (identified by Tax Identification Number [TIN]) and 8,087 practitioners (identified by combination of TIN and National Provider Identifier [NPI]). Clinicians and clinician groups are included if they are attributed 10 or more Routine Cataract Removal with Intraocular Lens (IOL) Implantation Measure (also referred to as "the Cataract measure") episodes, as identified in Medicare Parts A and B claims data, ending from January 1, 2017, to December 31, 2017. Episodes are included from all 50 States and D.C. in the following settings: ambulatory surgical center (ASC), ambulatory/office-based care, and hospital outpatient department (HOPD).

TIN Level Scores:

- Mean score: $3,041
- Standard deviation: $319
- Min score: $1,767
- Max score: $4,849
- Score IQR: $238
- Score percentiles
    - 10th: $2,794
    - 20th: $2,881
    - 30th: $2,922

    - 40th: $2,960
    - 50th: $3,002
    - 60th: $3,053
    - 70th: $3,106
    - 80th: $3,183
    - 90th: $3,378
- Number of beneficiaries: 490,714

TIN-NPI Level Scores

- Mean score: $3,038
- Standard deviation: $300
- Min score: $1,426
- Max score: $4,849
- Score IQR: $232
- Score percentiles
    - 10th: $2,811
    - 20th: $2,878
    - 30th: $2,915
    - 40th: $2,950
    - 50th: $2,992
    - 60th: $3,041
    - 70th: $3,097
    - 80th: $3,169
    - 90th: $3,363
- Number of beneficiaries: 485,216

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

N/A.

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

N/A.

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

N/A.

IM.2. Measure Intent

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

The Cataract measure was developed for use in the Merit-based Incentive Payment System (MIPS) in the Quality Payment Program (QPP), to meet the requirements of MACRA section 101(f). This program is required by law, and aims to achieve high-value care in the Medicare program by measuring clinician performance through four areas: quality, improvement activities, promoting interoperability, and cost. Within the MIPS cost performance category, this measure is intended to provide actionable information to clinicians about their cost

performance for cataract removals and IOL implantation, to allow them to make practical changes towards providing high-value, cost effective care.

Rationale for Measuring Cost through Episode-Based Cost Measure vs. All-Cost Measure

The intent of an episode-based cost measure is to capture only clinically related services within the reasonable influence of the attributed clinician, which is a key difference from broad, population-based cost measures such as the MIPS Total Per Capita Cost (TPCC) and Medicare Spending Per Beneficiary (MSPB) measures.

Episode-based cost measures represent the cost to Medicare for the items and services provided to a patient during an episode of care and are meant to inform attributed clinicians about the cost of care within their influence during the episode's timeframe. They represent a clinically cohesive set of medical services rendered to treat a given condition or related to a procedure; services are assigned to an episode only when clinically related to the attributed clinician's role in managing patient care during the episode.

Rationale for Measuring Cost of Routine Cataract Removal with IOL Implantation

Policymakers contend that an estimated 80 percent of overall health care costs are attributable to decisions made by clinicians.[1] 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 are significant opportunities for improvement, especially in mitigating costly complications as a result of the cataract surgery.

Cataract surgeries have a very high rate of surgical success with few postoperative complications as a result of technological advancements rendering it a much safer procedure. However, there are still some complications resulting from this procedure that may be mitigated through improved clinical practices, thereby resulting in cost savings from decreased downstream costs and fewer repeat surgeries.[2] A study found that one of the most frequent severe complication was endophthalmitis, a rare but serious eye infection complication, with an incidence between 0.06 and 0.20 percent among Medicare beneficiaries undergoing cataract surgery.[3] In a study assessing Medicare beneficiaries from 2010 to 2014, the adjusted Medicare claims and reimbursements for cataract surgery were 83 percent greater for beneficiaries that had developed endophthalmitis after the surgery. In addition, complications also arise that may result in a return to the operating room for follow up procedures. These are particularly significant as these returns add a sizeable additional cost on top of the initial cataract surgery.[4]

This measure aims to address these example areas of opportunities for improvement. Since cataract surgery is the most frequently performed surgical procedure in the United States, especially among Medicare beneficiaries, the use of this episode-based cost measure can provide clinicians with information to improve care outcomes and reduce future health care costs.

[1] 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.

[2] Pershing, S., D. E. Morrison, and T. Hernandez-Boussard. "Cataract Surgery Complications and Revisit Rates among Three States." [In eng]. Am J Ophthalmol 171 (Nov 2016): 130-38.

[3] Du, D. T., A. Wagoner, S. B. Barone, C. E. Zinderman, J. A. Kelman, T. E. Macurdy, R. A. Forshee, C. M. Worrall, and H. S. Izurieta. "Incidence of Endophthalmitis after Corneal Transplant or Cataract Surgery in a Medicare Population." [In eng]. Ophthalmology 121, no. 1 (Jan 2014): 290-8.

[4] French, D. D., C. E. Margo, and R. R. Campbell. "Comparison of Complication Rates in Veterans Receiving Cataract Surgery through the Veterans Health Administration and Medicare." [In eng]. Med Care 50, no. 7 (Jul 2012): 620-6.

Rationale for Use of Claims Data to Measure Cost

- The use of claims data for episode-based cost measures for MIPS is required by MACRA section 101(f).
- There is no additional submission burden, as clinicians must already submit claims for reimbursement.
- Using Medicare Parts A and B claims data allows CMS to evaluate TIN and TIN-NPI cost across all conditions and procedures, resulting in a comprehensive set of data on cataract removal and IOL implantation cost performance.
- Additionally, the wide reach of Medicare claims data maximizes the impact of the measure, ensuring that the most TINs and TIN-NPIs benefit from the information provided on cataract removal and IOL implantation cost performance.

Scientific Acceptability of Measure Properties

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

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

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

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

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

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

https://qpp.cms.gov/about/resource-library. Scroll to "Full Resource Library" to download the Cost Measure Information Form and Measure Code List, or see S.7.2a Construction Logic Attachment for additional details and specific links.

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

Per episode

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

Clinician : Group/Practice, Clinician : Individual

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

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

If other, please describe in S.5.1.

Claims

Enrollment Data

Other

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

The Routine Cataract Removal with IOL Implantation measure uses Medicare Part A and Part B claims data, which is maintained by CMS. Part A and B claims data are used to build episodes of care, calculate episode costs, and construct risk adjustors. Data from the Medicare Enrollment Database (EDB) are used to determine

beneficiary-level exclusions and supplemental risk adjustors, 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, and that information comes from the Minimum Data Set (MDS). 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 Common Medicare Enrolment (CME) are used in the analyses evaluating social risk factors in risk adjustment.

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

<SamplingMethodologySpecificDataSourceAttachment nodeType="0">S_5_2_DataSourceReference-636824776735244648.docx

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

Data Dictionary:

URL: The Research Data Assistance Center (ResDAC) maintains the Medicare claims data dictionary available here: http://www.resdac.org/cms-data/filefamily/Medicare-Claims CMS maintains the Medicare Enrollment Database and data dictionary: edbonline@cms.hhs.gov

Please supply the username and password:

Attachment:

Code Table:

URL:

Please supply the username and password:

Attachment: 2019_01_07_testing_form_appendix_cataract.xlsx

Construction Logic

S.7.1. Brief Description of Construction Logic

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

The Routine Cataract Removal with IOL Implantation measure is the sum of the ratio of observed to expected payment-standardized cost to Medicare averaged across the episodes attributed to a clinician or clinician group. This is then multiplied by the national average observed episode cost to generate a dollar figure. The measure can be calculated for an individual TIN-NPI (clinician) or a TIN (clinician group practice).

A Routine Cataract Removal with IOL Implantation episode is a unit or specific instance of the measure for a given clinician or clinician group and beneficiary that can then be aggregated to assess clinician performance across all their episodes. The episode is triggered or opened by a CPT/HCPCS code on Part B Physician/Supplier (Carrier) claims, and includes certain services in Medicare Parts A and B claims related to the procedure in the period 60 days before the episode trigger to 90 days after the episode trigger.

The cost measure numerator is the sum of the ratio of observed to expected payment-standardized cost to Medicare for all Routine Cataract Removal with IOL Implantation episodes attributed to a clinician or clinician group. Expected costs refer to costs predicted by the risk adjustment model. This sum is then multiplied by the national average observed episode cost to generate a dollar figure.

The cost measure denominator is the total number of episodes from the Routine Cataract Removal with IOL Implantation episode group attributed to a clinician or clinician group within a performance period (i.e., MIPS performance year).

Cost figures are standardized to remove the effect of differences in Medicare payment among health care providers that are the result of differences in regional health care provider expenses measured by hospital wage indexes, geographic price cost indexes (GPCIs), or other payment adjustments such as those for teaching hospitals. This standardization is intended to isolate cost differences that result from healthcare delivery choices, allowing for more accurate resource use comparisons between health care providers.

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

Step 1. Trigger and define an episode

Routine Cataract Removal with IOL Implantation episodes are defined by Current Procedural Terminology / Healthcare Common Procedure Coding System (CPT/HCPCS) codes on Part B Physician/Supplier (Carrier) claims that open, or trigger, an episode.

The steps for defining an episode for the Routine Cataract Removal with IOL Implantation episode group are as follows:

- Identify Part B Physician/Supplier claim lines with positive standardized payment that have a trigger code.
- Trigger an episode if all the following conditions are met for an identified Part B Physician/Supplier claim line:
    - It was billed by a clinician of a specialty that is eligible for MIPS.
    - It is the highest cost claim line across any Routine Cataract Removal with IOL Implantation trigger code billed for the beneficiary on that day.
    - It does not have a post-operative modifier code.[1]
- Establish the episode window as follows:
    - Establish the episode trigger date as the expense date of the trigger code.
    - Establish the episode start date as 60 days prior to the episode trigger date.
    - Establish the episode end date as 90 days after the episode trigger date.
- Define trigger exclusions based on information available at the time of the trigger, if applicable.

Once a Routine Cataract Removal with IOL Implantation episode is triggered, the episode is placed into one of the episode sub-groups to enable meaningful clinical comparisons. This cost measure has four sub-groups:

- ASC / Bilateral
- ASC / Unilateral
- HOPD / Bilateral
- HOPD / Unilateral

Step 2. Attribute the episode to a clinician

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

The steps for attributing a Routine Cataract Removal with IOL Implantation episode are as follows:

- Identify claim lines with positive standardized payment for any trigger codes that occur on the episode trigger day.
- Designate a TIN-NPI as a main clinician if the following conditions are met:
    - No assistant modifier code is found on one or more claim lines billed by the clinician.
    - No exclusion modifier code is found on the same claim line.

Designate a TIN-NPI as an assistant clinician if the following conditions are met:

  - The TIN-NPI was not designated as a main clinician.
  - An assistant modifier code is found.
  - No exclusion modifier code is found.

- Attribute an episode to any TIN-NPI designated as a main or assistant clinician.
- Attribute episodes to the TIN by aggregating all episodes attributed to NPIs that bill to that TIN. If the same episode is attributed to more than one NPI within a TIN, the episode is attributed only once to that TIN.

Step 3. Assign costs to the episode and calculate the episode observed cost

For the Routine Cataract Removal with IOL Implantation episode group, only services performed in the following service categories are considered for assignment to the episode costs:

- Outpatient (OP) Facility and Clinician Services

Service assignment rules may be modified based on the service category in which the service is performed, as listed above. Service assignment rules may also vary based on (i) additional criteria determined by other diagnosis, procedure, or billing codes appearing alongside the service code, or (ii) the specific timing of the service. Services may be assigned to the episode based on the following additional criteria:

- Services may be assigned to the episode based on the following additional criteria:
  - Service code alone
  - Service code in combination with other diagnosis, procedure, or billing codes such as:
    - The first three digits of the International Classification of Diseases – Tenth Revision diagnosis code (3-digit ICD-10 DGN)
    - The full ICD-10 DGN
    - Additional service information
- Services may be assigned only with specific timing:
  - Services may be assigned based on whether or not the service and/or diagnosis is newly occurring
  - Services may be assigned only if they occur within a particular number of days from the trigger within the episode window, and services may be assigned for a period shorter than the full duration of the episode window.

The steps for assigning costs are as follows:

- Identify all services on claims with positive standardized payment that occur within the episode window.
- Assign identified services to the episode based on the types of service assignment rules described above.
- Sum standardized Medicare allowed amounts for all claims assigned to each episode to obtain the standardized total observed episode cost.

Step 4. Exclude episodes:

The steps for episode exclusion are as follows:

- Exclude episodes from measure calculation if:
  - The beneficiary has a primary payer other than Medicare for any time overlapping the episode window or 120-day lookback period prior to the trigger day.

    - The beneficiary was not enrolled in Medicare Parts A and B for the entirety of the lookback period plus episode window, or was enrolled in Part C for any part of the lookback plus episode window.
    - No main clinician is attributed the episode.
    - The beneficiary's date of birth is missing.
    - The beneficiary's death date occurred before the episode ended.
    - The episode trigger claim was not performed in an OP hospital or ASC setting based on its place of service.
- Apply measure-specific exclusions, which check the beneficiary's Medicare claims history for certain billing codes (as specified in the Measure Codes List file) that indicate the presence of a particular procedure, condition, or characteristic.

Step 5. Calculate expected costs for risk adjustment:

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

- Define HCC and episode group-specific risk adjustors using service and diagnosis information found on the beneficiary's Medicare claims history in the 120-day period prior to the episode trigger day (or the timing specified in the "RA_Vars_Details" tab of the Measure Codes List file) for certain billing codes that indicate the presence of a procedure, condition, or characteristic.
- Define other risk adjustors that rely upon Medicare beneficiary enrollment and assessment data as follows:
    - Identify beneficiaries who are originally "Disabled without end-stage renal disease (ESRD)" or "Disabled with ESRD" using the original reason for joining Medicare field in the Medicare beneficiary enrollment database.
    - Identify beneficiaries with ESRD if their enrollment indicates ESRD coverage, ESRD dialysis, or kidney transplant in the Medicare beneficiary enrollment database in the lookback period.
    - Identify beneficiaries who have spent at least 90 days in a long-term care institution without having been discharged to the community for 14 days, based on MDS assessment data.
- Drop risk adjustors that are defined for less than 15 episodes nationally for each sub-group to avoid using very small samples.
- Categorize beneficiaries into age ranges using their date of birth information in the Medicare beneficiary enrollment database. If an age range has a cell count less than 15, collapse this with the next adjacent higher age range category.
- Run an ordinary least squares (OLS) regression model to estimate the relationship between all the risk adjustment variables and the dependent variable, the standardized observed episode cost, to obtain the risk-adjusted expected episode cost. A separate OLS regression is run for each episode sub-group nationally.
- Winsorize expected costs as follows [2].
    - Assign the value of the 0.5th percentile to all expected episode costs below the 0.5th percentile.
    - Renormalize values by multiplying each episode´s winsorized expected cost by the sub-group´s average expected cost, and dividing the resultant value by the sub-group´s average winsorized expected cost. [3]
- Exclude episodes with outliers as follows [4]. This step is performed separately for each sub-group.
    - Calculate each episode´s residual as the difference between the re-normalized, winsorized expected cost computed above and the observed cost.

- Exclude episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution.
- Renormalize the resultant expected cost values by multiplying each episode's winsorized expected costs after excluding outliers by the sub-group´s average standardized observed cost across all episodes originally in the risk adjustment model, and dividing by the sub-group´s average winsorized expected cost after excluding outliers.

Step 6. Calculate the measure score: Measure scores are calculated for a TIN or TIN-NPI as follows:

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

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

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

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

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

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

URL:

Please supply the username and password:

Attachment: S_7_2_Construction_Logic-636927567135380246.docx

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

To identify and manage concurrent clinical events, the measure takes into consideration the occurrence of the same cataract procedure on the other eye when it occurs within certain periods within the episode window. The measure provides specifications to account for a procedure that is performed on one side only (unilateral), both sides on the same day (same day bilateral), or both sides on different days within short succession (staged bilateral).

The episode construction methodology collapses two episodes together where there are staged bilateral procedures within 30 days; that is, where a beneficiary has a procedure on one side, then within 30 days (inclusive), the beneficiary has the same procedure on the other side. The rationale for this methodology is to capture the costs of assigned services which may not differentiated from available claims data as relating to the procedure on the left or right side – for example, a diagnostic imaging of retina service occurring during the

episode window could be for either eye. By collapsing the two episodes together, the measure captures the overall costs of care for the two closely related procedures occurring within a short span of time.

The risk adjustment methodology accounts for laterality in the following way:

- Establish a unilateral sub-group if the procedure is performed for one eye during the trigger event.
- Establish a bilateral sub-group if one of the following conditions is true:
    - The procedure is performed for both eyes during the trigger event; or
    - A procedure on one side (identified by the sole presence of right [or left] modifier code) is followed by the opposite side procedure within 30 days. In these staged bilateral events, the two episodes will be combined to one.

Using this methodology to establish sub-groups, the measure accounts for differences in episode cost for unilateral and bilateral procedures. The rationale for having sub-groups based on unilateral and bilateral surgery is to account for some services (e.g., preoperative exams and testing) that may be applied to a second surgery performed in close succession.

This measure is designed to allow episodes to overlap with other episodes: overlapping episodes are different episodes that are triggered for the same patient with overlapping episode windows. The advantage of this is that each episode can reflect attributed clinicians' different roles in providing care services throughout a patient's care trajectory. For example, a patient could have a Routine Cataract Removal with IOL Implantation episode triggered when the attributed clinician performs the procedure, and 80 days later be admitted to hospital for pneumonia unrelated to the cataract surgery, triggering an episode for a different cost measure that is attributed to the hospitalist providing care for pneumonia. Each episode includes only the cost of assigned services (i.e., those that are within the reasonable influence of the attributed clinician) to reflect each attributed clinician's role. In addition, costs are not double counted as the measure calculation is based on the ratio of observed over expected spending for each episode, then averaged across all of an attributed clinician's episodes.

The measure accounts for disease interactions through its risk adjustment model based on the CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) 2016 model. In addition to the HCCs, the model includes disease interactions (e.g., Cancer * Immune Disorders). Further details about the risk adjustment model and disease interaction terms are included in Section S.8.6.

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

This measure includes the cost of services that are clinically related to the procedure for routine cataract removal with IOL implantation. The rationale for only including specific costs is to ensure that the attributed clinician is evaluated only on his or her performance on services over which they have reasonable influence. For instance, the cost of anesthesia for lens surgery is included in a clinician's episode cost if it occurs any time during the episode window.

The assigned services for this measure have been identified as being related to the procedure and within the influence of the attributed clinician through consideration of detailed input from clinician experts and broader feedback from stakeholders from the clinician community. Specifically, an Ophthalmologic Disease Management Clinical Subcommittee was convened from May 2017 to January 2018 to discuss and provide detailed recommendations on aspects of measure construction, including the services to be included in this measure. This Subcommittee was composed of 10 clinician experts affiliated with 11 specialty societies.

Members reviewed analyses of the utilization and timing of all Medicare Parts A and B services in broad timeframes extending before and after the episode trigger to provide recommendations on the services and associated conditions for including these as part of the episode costs. For example, the subcommittee decided that the cost of hospitalization is not relevant to the episode therefore is not assigned to episode. Conditions could include requiring additional codes to be present on services to ensure clinical relevance, assigning for a shorter timeframe within the overall episode window, or assigning only if a diagnosis that is part of the trigger

logic is newly occurring. The draft measure was field tested from October to November 2017: during this time, stakeholders reviewed the measure specifications, including a list of assigned services and associated logic conditions, field test reports containing details of attributed clinician performance, and supplemental documentation. Over 65,000 TIN and TIN-NPI field test reports were available during this time for review and feedback.

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

National Summary Data Report (July 2018) – this document contains summary data about the Routine Cataract Removal with IOL Implantation cost measure, along with other episode-based cost measures. These summary statistics supplement the testing analyses contained in this submission: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2017-field-test-materials.zip, filename: 2018-07-12-national-summary-data-report.pdf

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

Field Testing Feedback Summary Report (June 2018) – this document summarizes the feedback received during a stakeholder feedback period during measure development. The Routine Cataract Removal with IOL Implantation cost measure has been developed with extensive input from the clinician community: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2018-field-testing-feedback-summary-report.pdf

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

Clinical hierarchies are embedded in the risk adjustment model. The risk adjustment model includes variables from the CMS-HCC V22 2016 Risk Adjustment Model, as well as other standard risk adjustors (e.g., beneficiary age brackets using information in the Medicare beneficiary enrollment database) and disease interaction terms. The model also includes variables specific to this cost measure, identified through the incorporation of detailed clinical input. These variables account for whether a patient is a new or established patient and whether the surgery is performed by a resident under the direction of a teaching physician.

The CMS-HCC V22 model uses 79 Hierarchical Condition Category (HCC) indicators derived from the beneficiary's claims in the period 120 days prior to the episode trigger day. Other risk adjustors are originally "Disabled without end-stage renal disease (ESRD)" or "Disabled with ESRD" using the original reason for joining Medicare in the Medicare beneficiary enrollment database. The risk adjustment model also uses an indicator for beneficiaries identified as having had recent need of long-term care (90 days in a long-term care institution without having been discharged to community for 14 days) using MDS assessment data. Additional information about the risk adjustment model is included in Section S.8.6.

The Routine Cataract Removal with IOL Implantation episode group includes all services identified as being clinically relevant to this procedure and within the reasonable influence of the attributed clinician. There are logic rules to determine when and what conditions each particular service will be assigned, as detailed in the Measure Codes List file (see Section S.1 for URL).

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

All the data used to calculate the Routine Cataract Removal with IOL Implantation cost measure are included on Medicare claims data. The data fields used to calculate measure (e.g., payment amounts, diagnosis and procedure codes, etc.) are included in all Medicare claims because clinicians only receive payments for complete claims. Additional information regarding the reliability of diagnostic information on claims is available on the Testing Form in Section 2a2.2.

We have complete data for each beneficiary who opens an episode by receiving a triggering service, since beneficiaries are excluded if they are not continuously enrolled in only Medicare Parts A and B or if Medicare is not the primary payer during an episode. This ensures that we have all claims data for beneficiaries included in the Routine Cataract Removal with IOL Implantation cost measure.

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

Ambulatory services: Outpatient facility services

Ambulatory services: Evaluation and management

Ambulatory services: Procedures and surgeries

Ambulatory services: Imaging and diagnostic

Ambulatory services: Lab services

Other ambulatory services

See Measure Codes List

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

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

The Routine Cataract Removal with IOL Implantation measure assesses the standardized allowed amounts of services performed by clinicians and other healthcare providers during an episode, which includes all assigned services from Part A and Part B Medicare claims that occur within the time period 60 days prior to the episode trigger through 90 days after the trigger.

The assigned services for this measure are within the outpatient facility and clinician services category. The CPT/HCPCS codes to identify these services are contained in the Measure Codes List file (see Section S.1), along with the logic conditions for assigning these services.

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

URL:

Please supply the username and password:

Attachment:

Clinical Logic

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

The measure aims to provide actionable information to clinicians performing a routine cataract removal with IOL implantation procedure about their resource use within the overall goal of enabling clinicians to provide cost-effective and high-quality care. The clinical logic is constructed to achieve this objective.

Clinical Topic Area: Cataract removal

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

Clinical Hierarchies: Clinical hierarchies are embedded in the risk adjustment model. See Section S.7.5. for further details.

Clinical Severity Levels: The measure has sub-groups to account for bilateral and unilateral procedures, and excludes patients with ocular comorbidities as part of the measure intent of focusing on routine cataract removal procedures.

Concurrency of Clinical Events: The measure spans the period from 60 days prior to the episode trigger to 90 days after the episode trigger. Services that are clinically related to the procedure and within the reasonable influence of the attributed clinician within this period of time are included in the episode. The measure accounts for unilateral and bilateral procedures. See Section S.7.3. and S.7.4. for further details.

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

The Routine Cataract Removal with IOL Implantation measure uses a service assignment algorithm and includes services for cataract surgery evaluation, testing, treatment, complications, and follow-up.

Grouping methodology and assignment algorithm: The Routine Cataract Removal with IOL Implantation cost measure evaluates resource use through the unit of episodes of care. The cost measure episodes are constructed by including select Medicare Part A and Part B claims (assigned services) which occur during the episode window, defined as 60 days prior to the episode trigger to 90 days after the trigger. The episode trigger and assigned services are contained in the Measure Codes List file (see Section S.1. for details), along with risk adjustors, sub-groups, and exclusions.

Details about the measure exclusions are in Section S.9.1.

Cost Calculation: The cost measure amount includes the cost of assigned services performed by clinicians and other providers during the episode window. The cost measure is calculated as the sum of the ratios of observed to expected costs, multiplied by the national average observed episode cost to generate a dollar figure, and then divided by total number of episodes from the episode group attributed to a clinician. All costs are payment standardized to control for geographic variation in Medicare reimbursement rates. The measure is risk adjusted to account for age and severity of illness. Expected costs are estimated through risk adjustment by using an ordinary least squares regression model. More details about the risk adjustment model are described in Section S.8.6.

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

The clinical logic used in the Routine Cataract Removal with IOL Implantation measure is informed by literature review, stakeholder feedback, and clinician input.

A study notes that policymakers contend that an estimated 80 percent of overall health care costs are attributable to decisions made by clinicians (Fred, 2016). However, these same clinicians are often unaware of how their care decisions influence the overall costs of care. One of the goals of the use of cost measures is to help inform clinicians on the costs for which they are directly responsible, as well as the total cost of their patient's care. A cost measure exhibits the 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 made through changes in clinical practice.

Since cataract surgery is the most frequently performed surgical procedure in the United States, especially among Medicare beneficiaries, the use of this cost measure can provide clinicians with information to improve care outcomes and reduce future health care costs. Cataract surgeries have a very high success rate with few postoperative complications, as technological advancements have rendered it a much safer procedure. However, there are still some complications resulting from this procedure that may be mitigated through improved clinical practices, thereby resulting in cost savings from decreased downstream costs and fewer repeat surgeries (Pershing et al., 2016).

The measure was designed to incorporate extensive expert clinician input into each component of the measure to ensure that it achieves the goal of providing actionable information to clinicians for their performance of a procedure on a coherent patient cohort. The measure was developed to meet the requirements of MACRA section 101(f) to create episode-based cost measures. It aligns with CMS meaningful measure area of 'patient-focused episode of care' within the overall quality priority of 'Make Care Affordable'.

The measure includes services that are clinically related to the procedure and within the reasonable influence of the attributed clinician. By including services after the procedure, it aims to improve care coordination throughout a patient's care trajectory. The measure also aligns with quality measures for the same clinical area by focusing on the same patient cohort who undergo this procedure without significant ocular conditions (see NQF #0564 and NQF #0565).

Fred, H. L. "Cutting the Cost of Health Care: The Physician´s Role." [In eng]. Tex Heart Inst J 43, no. 1 (Feb 2016): 4-6

Pershing, S., D. E. Morrison, and T. Hernandez-Boussard. "Cataract Surgery Complications and Revisit Rates among Three States." [In eng]. Am J Ophthalmol 171 (Nov 2016): 130-38.

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

URL:

Please supply the username and password:

Attachment: 2018-12-21-codes-list-cataract-636927598904672528.xlsx

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

The detailed steps for triggering Routine Cataract Removal with IOL Implantation episodes are in Section S.7.2. The advantage of the simplicity in opening episodes this way is to ensure that clinicians know at the time of providing the service that an episode has been triggered. This helps meet the goal of the measure to provide actionable information to clinicians.

Additional conditions must be met to trigger an episode. The triggering procedure must take place in an outpatient hospital setting or ASC.

The Routine Cataract Removal with IOL Implantation episode window is defined as follows:

- Episode trigger date: expense date of trigger code
- Episode start date: 60 days prior to episode trigger day
- Episode end date: 90 days after episode trigger date

As discussed in Section S.7.3, staged bilateral episodes occurring within 30 days of the initial procedure on the other side are collapsed into one episode. This means that the trigger date will be the date of the first procedure (on one side) and the episode end date will be 90 days after the second procedure (on the other side). This ensures that assigned services for both procedures are captured.

The conditions to trigger episodes and the duration of the episode window were established with input from clinician experts in consideration of the goals of the measure to provide actionable information to clinicians about their resource use for a comparable patient cohort. An initial Draft List of Episode Groups and Trigger Codes was posted in December 2016 incorporating input from a Clinical Committee of more than 70 clinicians from over 50 professional societies. Feedback from a four-month public comment period on that posting was summarized and shared with the Ophthalmologic Disease Management Clinical Subcommittee who used the information from the draft list as a starting point and took feedback into consideration along with analyses to help inform discussions, such as the frequency of services over a period of time extending from the trigger date. The pre-trigger window was informed by data indicating the presence of a clinic visit for 77 percent of episodes within 60 days pre-trigger. The post-trigger window was determined for consistency with the postoperative 90-day global period. This measure was field tested in 2017, as discussed further in Section S.7.4. The Clinical Subcommittee took field testing feedback into consideration in making refinements to the measure, including feedback on episode triggers and episode window length.

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

Clinical severity levels are embedded in the risk adjustment model, as described in Section S.7.5.

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

The Routine Cataract Removal with IOL Implantation measure accounts for comorbid conditions and interactions by broadly following the CMS-HCC risk-adjustment methodology, which is derived from Medicare Part A and B claims and is used in the Medicare Advantage (MA) program. Diagnosis codes on claims that occur during the 120-day period prior to the episode trigger date are used to create HCC indicators. Episodes where the beneficiary is not enrolled in both Medicare Part A and Medicare Part B for the 120 days prior to the episode are excluded because information on comorbidities for these beneficiaries will be incomplete. When applying the CMS-HCC framework to the measure, expected costs are determined by the risk adjustment model separately for each sub-group, which allows the effect of beneficiary health status and demographics on episode spending levels to vary by the sub-groups which reflect place of service and laterality. This cost measure accounts for comorbid interactions by incorporating a number of health status interactions as currently used within the CMS-HCC model. The model includes paired-condition interactions (e.g., chronic obstructive pulmonary disease (COPD) and congestive heart failure (CHF)) and interactions between conditions and disability status (e.g., disabled and cystic fibrosis). There are also variables for whether it is a new or established patient, and where the surgery is performed by a resident under the direction of a teaching physician. The full list of variables used in the risk adjustment model can be found in the Measure Codes List, linked at Section S.1.

The 120-day period prior to the start of an episode is used to identify the conditions which most directly impact beneficiaries' health status at the time of the procedure and to capture beneficiaries' comorbidities in the risk adjustment. Additionally, because the relationship between comorbidities' episode cost may be non-linear in some cases (i.e., beneficiaries may also have more than one

disease during a hospitalization episode), the model also takes into account a limited set of interactions between HCCs and/or enrollment status variables. The Routine Cataract Removal with IOL Implantation measure risk adjustment methodology includes only a limited set of interaction terms for two reasons. First, inclusion of too many interaction terms will over-fit the model. Second, the risk-adjustment methodology broadly follows the established CMS-HCC risk-adjustment methodology, which uses similar interaction terms.

Adjustments for Comparability

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

Included populations:

The beneficiary population eligible for the Routine Cataract Removal with IOL Implantation measure calculation consists of Medicare beneficiaries enrolled in Medicare Parts A and B who received a cataract removal procedure during the performance period as identified by the episode trigger code. To be included, the beneficiary must have an episode ending within the performance period to ensure that the beneficiary's claims record contains sufficient fee-for-service data both for measuring spending and for risk adjustment purposes.

Excluded populations:

Episodes are excluded for the following conditions, with the rationale for each provided below.

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

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

- No attributed clinician is found for the episode:

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

- The beneficiary's date of birth is missing:

These episodes are excluded as a data cleaning step.

- The beneficiary's death date occurred before the trigger date:

These episodes are excluded as a data cleaning step.

- The beneficiary's death date occurred before the episode ended:

Episodes ending in death are excluded as they are - by definition - truncated episodes and do not have a complete episode window. Including episodes without all observable claims or a complete episode window could potentially make clinicians appear to have lower cost episodes not due to efficiencies of their own performance, but because the data are missing services that would be included in the measure calculation.

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

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

- The episode trigger claim was not performed in an outpatient hospital or ASC setting

Episodes where the Part B Physician/Supplier claim with the CPT/HCPCS trigger code is not performed in an outpatient hospital or ASC are excluded to ensure that this measure captures a homogenous patient cohort, focusing on uncomplicated cataract removal procedures. Performing this procedure in other settings could indicate more complex procedures.

- Episodes where the beneficiary has ocular comorbidities (impacting visual outcome of surgery or surgical complication rate)

Beneficiaries with significant ocular conditions are excluded from this measure, as defined by the presence of ICD-10 diagnosis and CPT/HCPCS codes on Part B Physician/Supplier, Outpatient, and Inpatient claims in a 120-day lookback period. These diagnosis and procedure codes indicate significant ocular conditions (e.g., diabetic retinopathy) that impact the outcomes of surgery and expected resource use. Patients with these conditions are more likely to require more complex care that differs from the routine care in this measure. Exclusion of these patients is consistent with MIPS quality measures assessing the outcome of routine cataract surgery (NQF #0564 and #0565).

- 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.

S.9.2. Risk Adjustment Type (Select type)

Stratification by risk category/subgroup

If other:

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

The Routine Cataract Removal with IOL Implantation measure is stratified into four sub-groups: ASC/Bilateral, ASC/Unilateral, HOPD/Bilateral, and HOPD/Unilateral. The stratification for site of service accounts for access factors, as some clinicians may not have access to an ASC, a lower cost setting than HOPD, due to regional availability or as a result of health plan contracting arrangements. Sub-groups for unilateral and bilateral surgery are used to account for scenarios where some services may be applied to a second surgery performed in close succession, meaning that bilateral procedures will likely be more expensive than unilateral ones. These sub-groups represent more homogenous patient cohorts to enable meaningful clinical comparisons based on information available on the trigger claim. These sub-groups are useful in ensuring clinical comparability so that the corresponding cost measure fairly compares clinicians with a similar patient case-mix. A separate risk adjustment model is created for each stratified group, so that clinically meaningful distinctions in the beneficiary population are preserved.

S.9.4 Costing method

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

Standardized pricing

The methodology used to payment standardize the Medicare claims used to specify this measure is available for download ("CMS Price (Payment) Standardization") from the URL provided below.

http://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890990237&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DDetailed_Mthds_payment-std_041819.pdf&blobcol=urldata&blobtable=MungoBlobs

This direct-download link changes biannually as the documentation is updated; if the link no longer works, the download may also be accessed via the page linked below.

https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1228772057350

Detailed_Mthds_payment-std_041819-636927619044855059.pdf

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

Ratio

If other:

Attachment:

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

The Routine Cataract Removal with IOL Implantation cost measure score is a dollar value that represents a clinician's average payment-standardized risk-adjusted cost to Medicare across all Routine Cataract Removal with IOL Implantation episodes attributed to them. A value above the national average indicates that on average, the clinician's resource use for this procedure was more expensive than the national average. A value below the national average indicates that on average, the clinician's resource use for this procedure was less expensive than the national average.

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

quality measures. The focus of this measure is on patients receiving uncomplicated cataract removal procedures, and excludes patients with significant ocular conditions, which aligns with the focus of two NQF-endorsed quality measures (see NQF #0564 and NQF #0565).

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

A clinician's Routine Cataract Removal with IOL Implantation measure score is calculated as the average ratio of observed cost to expected episode cost across a provider's episodes, multiplied by the national average observed episode cost. This calculation is done using episodes from all sub-groups. Further details are provided in Section S.7.2.

Reporting Guidelines

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

S.13.1. Describe discriminating results approach

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

The measure is used in MIPS for the CY 2019 performance period onwards. As such, it has not yet been reported as part of MIPS scoring. However, during measure development, we conducted national field testing where confidential reports containing cost measure performance on the Routine Cataract Removal with IOL Implantation measure at its draft stage of development (and other episode-based cost measures developed at the same time) were available to clinicians and clinician groups meeting a 10-episode case minimum. The purpose of this field testing was to enable clinicians to become familiar with the measure and to provide feedback on the measure specifications for refinement before CMS considered the measure for use in MIPS. During field testing, a National Summary Data Report was also posted containing summary statistics on the episode-based cost measures, including information on the distribution of TIN and TIN-NPI level measure scores.

S.13.2. Detail attribution approach

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

The Routine Cataract Removal with IOL Implantation episode is attributed to clinicians (TIN-NPIs) billing the episode trigger code. The episode is attributed to a TIN by aggregating all episodes attributed to the NPIs that bill to that TIN. If the same episode is attributed to more than one NPI within a TIN, this episode is only attributed to the TIN once. This allows the measure to be reported to both TINs and TIN-NPIs.

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

S.13.3. Identify and define peer group

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

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

S.13.4. Sample size

Detail the sample size requirements for reporting measure results.

The Routine Cataract Removal with IOL Implantation measure will be reported for TINs and TIN-NPIs with 10 or more episodes. The measure is used in the Merit-based Incentive Payment System (MIPS) for MIPS performance period 2019 onwards.

S.13.5. Define benchmarking and comparative estimates

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

The measure has not been reported yet, as it will be used in the MIPS cost performance category for the 2019 performance period onwards.

Reporting this measure as part of the cost performance category helps to measure clinicians' resource use for the routine cataract removal with IOL implantation procedure in the Medicare population, and thereby hold clinicians accountable for their cost effectiveness. There is no reporting/data submission requirement. Combined with measures in the other MIPS performance categories, such as the quality performance category, the Routine Cataract Removal with IOL Implantation measure allows CMS to assess the value of care and incentivize both achievement and improvement in the provision of high-quality, cost-effective care.

Validity – See attached Measure Testing Submission Form

SA.1. Attach measure testing form

2019_04_16_nqf_testing_form_cataract.docx

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** N/A
Measure Title: Routine Cataract Removal with IOL Implantation
Date of Submission: 4/16/2019

Type of Measure:

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

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

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

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

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

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Medicare Parts A and B claims data from the Common Working File (CWF); Long-term Minimum Data Set (MDS) data; Enrollment Database (EDB) data; Common Medicare Environment (CME); and the United States Census Bureau's American Community Survey (ACS).

1.3. What are the dates of the data used in testing? Routine Cataract Removal with IOL Implantation episodes ending from January 1, 2017 to December 31, 2017. For further details, please see Question 1.7.

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

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

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

There were 4,515 clinician group practices (identified by Tax Identification Number [TIN]) and 8,087 practitioners (identified by combination of TIN and National Provider Identifier [NPI]) included in the analysis. Clinicians and clinician groups were included if they were attributed 10 or more Routine Cataract Removal with IOL Implantation (also referred to as "the Cataract measure") episodes, as identified in Medicare Parts A and B claims data, ending from January 1, 2017, to December 31, 2017. Episodes were included from all 50 States and D.C. in the following settings: ambulatory surgical center (ASC), and hospital outpatient department (HOPD).

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

There were 490,714 Medicare beneficiaries (from 555,165 episodes) included in TIN level testing and analysis, and 485,216 beneficiaries (from 548,945 episodes) included at TIN-NPI level measure testing. Routine Cataract Removal with IOL Implantation episodes are triggered by a Current Procedural Terminology (CPT) / Healthcare Common Procedure Coding System (HCPCS) on Part B Physician/Supplier claims which indicates occurrence of a routine cataract removal procedure. 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 routine cataract removal procedures. As previously mentioned, a 10 episode case minimum was also applied. These inclusion criteria are listed below:

- 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 episode can be attributed to at least one main clinician.
- The episode trigger claim was in an HOPD or ASC setting based on its place of service.
- The beneficiary does not have significant ocular comorbidities.
- The episode is not an outlier case.

To determine whether the Cataract 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.

Appendix Table 1.6 details these distributions and shows that the Cataract measure inclusion criteria have only a minimal effect on the percentage of beneficiaries of any particular demographic. The difference between beneficiaries being included or not included in the measure is less than 0.4 percentage points across each of the characteristics in the analysis at TIN level testing, and less than 0.2 percentage points at TIN-NPI level testing. To illustrate, the percentage of beneficiaries aged 65 to 69 without applying the inclusion criteria is 25.6 percent, compared to 25.4 percent when the inclusion criteria are applied at TIN level testing, and remains at 25.7 percent at TIN-NPI level testing. The difference in the percentage of beneficiaries for race with and without the inclusion criteria is 0.0 percentage points for most categories, and is between 0.1 and 0.2 percentage points for two categories (Black and White) for TIN and TIN-NPI testing. The breakdown of male and female beneficiaries remains the same when comparing the use of inclusion criteria at TIN-NPI level testing, with 62.2 percent female and 37.8 percent male either with or without the application of inclusion criteria. At TIN level testing, there is a difference of 0.4 percentage points between the share of male and female beneficiaries when comparing the effect of inclusion criteria. These results indicate that there is minimal shift in patient characteristics as a results of using the inclusion criteria listed above at both TIN and TIN-NPI level testing.

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

Measure reliability scores for Section 2a2.3 are taken from the publically available National Summary Data Report on Eight Wave 1 Episode-Based Cost Measures.[1] These scores were calculated using episodes ending from June 1, 2016 to May 31, 2017.

All other testing used the study period outlined above in Section 1.3 from January 1, 2017 to December 31, 2017.

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

[1] Centers for Medicare & Medicaid Services. "National Summary Data Report on Eight Wave 1 Episode-Based Cost Measures: Elective Outpatient Percutaneous Coronary Intervention (PCI), Knee Arthroplasty, Revascularization for Lower Extremity Chronic Critical Limb Ischemia, Routine Cataract Removal with Intraocular Lens (IOL) Implantation, Screening/Surveillance Colonoscopy, Intracranial Hemorrhage or Cerebral Infarction, Simple Pneumonia with Hospitalization, ST-Elevation Myocardial Infarction (STEMI) with PCI", July 2018. Table 2. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2017-field-test-materials.zip

The social risk factors analyzed were variables from the American Community Survey (ACS), Medicare Enrollment Database (EDB), and Common Medicare Enrollment (CME). Please note that all ACS variables are at the Census Block Group level. Social risk variables analyzed include the following:

1. Income (ACS): Low Income (when median income < 33rd percentile nationally); Medium Income (when median income in the interval spanning the 33rd percentile to the 66th percentile nationally); High Income (when median income > 66th percentile).
2. Education (ACS): Education < High School (when % with < high school education is the highest for a given Census Block Group); Education = High School (when % with only high school is the highest); Education > High School (when % with > high school is the highest).
3. Employment (ACS): Unemployment Rate > 10%; Unemployment Rate <= 10%.
4. Race (EDB): Asian, Black, Hispanic, North American Native, White, and Other
5. Sex (EDB): Female, Male
6. Dual status (CME): Full Dual, Partial Dual, Non-dual.

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

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

Data Element Reliability

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

CMS also uses the Comprehensive Error Rate Testing (CERT) Program to ensure that Medicare payments are correct in accordance with coverage, coding, and billing rules. Between 2005 and 2017, CERT estimates that proper payment, which is payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments each year.[2] The FY 2018 Medicare FFS program proper payment rate was 91.9 percent.[3] 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 calculated using data with a three month claims run-out from the end of the performance period.

---

[2] 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/2018MedicareFFSSuplementalImproperPaymentData.pdf

[3] Ibid.

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 utilize two approaches: (1) Test/Retest and (2) Reliability Score.

Our first approach to assess reliability is to consider the extent to which assessments of a clinician using unique sets of episodes produce similar measures of clinician performance. That is, we take a "test-retest" approach in which performance is measured using two sets of episodes. We then examine the correlation and quintile rank stability between a TIN or TIN-NPI's cost measure scores calculated from both samples. Specifically, we ranked clinicians by their score within each sample and stratified clinicians into quintiles (with Quintile 1 representing the lowest cost and Quintile 5 the highest). We then calculated the percentage of clinicians who changed in measure score quintile between the two samples.

By comparing the scores of each TIN and TIN-NPI calculated using the two mutually exclusive samples, one can identify how consistently the measure identifies clinician performance. For this analysis, Acumen compared two random sets of episodes to identify the reliability of TIN or TIN-NPI score across samples.

Our second 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 and σ_b^2 is the between-group variance of clinician 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.

The following section provides a clarification of the interpretation of the test/retest results. While the reliability testing for this measure was not discussed at the Scientific Methods Panel meetings, the subgroup referenced Adams (2010) in the discussion of the test/retest results for the reliability of two other episode-based cost measures submitted within the same cycle. Since the test/retest analysis had also been included for this measure, we would like to clarify that this analysis cannot be interpreted as akin to the analysis conducted in Adams (2010).

First, each of the test/retest split samples has fewer cases per provider than a full-year performance period would have. This systematically results in reduced precision relative to how the measure would be used in practice. Consequently, the test-retest analyses likely understate the stability in provider measure scores. Second, the test/retest analysis is conceptually distinct from the type of analysis in Adams (2010). The Adams type of reliability method must first take some number (e.g., the provider's observed mean score) as the assumed value for a provider's performance and test the influence of "statistical noise" on provider classification. This statistical noise is given by the "within variance" of the mean score, which is the σ_w^2 term used in the signal-to-noise reliability equation (above). By contrast, test/retest split-sample quintile shifts simply consider both sample scores to be two noisy estimates of the provider's underlying score. Since this is a conceptually distinct exercise, equating a test/retest analysis with a reliability analysis of the type pursued in the Adams paper yields a misleading interpretation.

Change in Provider Classification (Adams)

To address the SMP's expressed interest in Adams (2010), we provide an additional analysis that aims to investigate the relationship between the provider's measured score on this cost measure and their true

performance relative to other providers. This analysis uses the variance in each provider's performance to calculate how often a provider's measured score is in the same performance tier as that of their true score.

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

Test/Retest

Across the two random samples of episodes (with a 10 episode case minimum applied for each sample), we see a Pearson correlation of 0.93 at the TIN level and 0.92 at the TIN-NPI level, and limited movement across quintiles. Over 70 percent of TINs and 67 percent of TIN-NPIs in the lowest-spending quintile (Quintile 1) in the first sample are also in the lowest-spending quintile in the second sample. Moreover, more than 91 percent of TINs and 89 percent of TIN-NPIs in the lowest-spending quintile in the first sample are in one of the two lowest spending quintiles (Quintiles 1 and 2) in the second sample.

Similarly, nearly 79 percent of TINs and over 76 percent of TIN-NPIs are in the highest-spending quintile (Quintile 5) in both samples, with approximately 94 percent of TINs and TIN-NPIs in the highest-spending quintile in one sample and in the top two highest quintiles (Quintiles 4 and 5) in the next. Full quintiles results are listed in Appendix Table 2a2.3.

Reliability Score

Using the methodology outlined in the previous section, Acumen previously calculated reliability scores that are publically available in the National Summary Data Report on Eight Wave 1 Episode-Based Cost Measures.[4] These scores were calculated using episodes ending from June 1, 2016 to May 31, 2017.

Based on these scores, 100 percent of TINs and TIN-NPIs have a reliability score greater than 0.4, the standard that CMS generally considers as the threshold for 'moderate' reliability. With a 10 episode case minimum, the mean reliability for TINs is 0.95 and for TIN-NPIs is 0.94.

The full reliability distribution at the listed case minimum is as follows. To address the SMP's interest in seeing reliability at varying case minimum thresholds for another episode-based cost measure, we have also provided reliability at a 20 and 30 episode case minimum. Based in part on these results, the measure as used in the MIPS CY 2019 performance period uses a case minimum of 10 episodes. A higher volume threshold or case minimum would have yielded even higher reliability, but at the cost of further reducing the number of clinicians and clinician groups able to receive a score.

Reporting Level	Case Minimum	Mean Reliability	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile
TIN	10	0.952	0.880	0.932	0.966	0.985	0.993
TIN-NPI	10	0.940	0.869	0.917	0.955	0.975	0.985
TIN	20	0.966	0.928	0.949	0.973	0.987	0.993
TIN-NPI	20	0.958	0.917	0.941	0.963	0.978	0.986
TIN	30	0.973	0.944	0.960	0.977	0.988	0.994
TIN-NPI	30	0.966	0.939	0.953	0.968	0.980	0.987

Change in Provider Classification (Adams)

We also analyze reliability in a manner analogous to the Adams paper. We classify each provider into "Low Cost" if below the 25th percentile of provider's scores and "Not Low Cost" if above. Using methods similar to

[4] CMS. "National Summary Data Report", July 2018. Table 2. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2017-field-test-materials.zip

the reliability calculation, we then computed each provider's standard error of the mean score. We then estimate the probability of a provider shifting to a different classification, and take the mean across these probabilities. For a measure with high stability, we would expect only a small proportion of low cost providers to be categorized as high cost (or vice-versa). However under this methodology, due to the existence of some clinicians very close to the border (e.g. within $5 of the low cost classification), we would expect some movement in classification. The results of this analysis show that an average provider has a 9.7% probability of being classified as "Low Cost" when they should be classified as "Not Low Cost" or vice-versa.

For illustrative purposes, given the SMP's interest, we used similar techniques to provide a quintile analysis that is more analogous to the Adams method than the test-retest analyses above. However, while this analysis using quintiles is more similar to the Adams paper, it is important to note a key distinction: by definition, an analysis of scores using more tiers will have greater movement than an analysis using fewer classes, due to the movement of scores for providers near the cutoffs. As such, using five quantiles (such as in the analysis below) rather than only two in the Adams paper will see greater movement. Acumen calculated the probability that a provider with performance in one quintile is classified to a different quintile, and averaged these probabilities across providers. Based on this analysis, 93% of the time, providers with a measure score in the lowest-spending quintile (Quintile 1) will remain in that quintile. Similarly, 99% of the time, providers with a measure score in the highest-spending quintile (Quintile 5) will remain in that quintile.

TIN Performance Quintile	Probability of Measure Score in Quintile				
	Q1	Q2	Q3	Q4	Q5
Quintile 1	93.4%	4.6%	1.3%	0.5%	0.2%
Quintile 2	17.0%	66.2%	13.7%	2.5%	0.6%
Quintile 3	3.7%	12.5%	70.5%	11.7%	1.6%
Quintile 4	0.9%	1.6%	9.6%	79.8%	8.1%
Quintile 5	0.1%	0.1%	0.2%	1.0%	98.7%

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Test/Retest

The test/retest results indicate the Cataract measure is robust to statistical noise resulting from random sampling and is therefore expected to consistently reproduce the same result. More specifically, the test/retest analysis shows that a large majority of clinicians have similar measure quintile ranks regardless of which episodes are used to calculate this scores. This indicates that the Cataract measure score is a reliable measure of a clinician's risk-adjusted Medicare spending compared to other clinicians.

As noted previously, the requirement to split the measure into two samples results in each provider having fewer cases than in a full performance period. This systematically results in reduced precision relative to how the measure would be used in practice. Consequently, per Adams, the test-retest analyses can only be considered a lower bound on the stability in provider measure scores[5]. The quintile tables for the Change in Provider Classification analysis therefore provide a better estimate of provider stability.

Reliability Score

[5] Adams, John L., The Reliability of Provider Profiling: A Tutorial. Santa Monica, CA: RAND Corporation, 2009. https://www.rand.org/pubs/technical_reports/TR653.html

Overall reliability of the Cataract measure is very high at a 10-episode case minimum for both TINs and 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.[6] Applying a case minimum of 10 episodes per clinician or clinician group ensures both high reliability and measure inclusiveness.

Change in Provider Classification (Adams)

The analysis results indicate that the percentage of low cost providers classified as high cost, or vice versa, is low. More specifically, this analysis indicates that measure score rankings are largely driven by differences in true clinician performance, and that random noise has little effect on the results of the measure.

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Face validity

The Cataract measure was developed with extensive input from clinician experts and other stakeholders. Acumen incorporated input from (i) an Ophthalmologic Disease Management Clinical Subcommittee (CS), (ii) Technical Expert Panel (TEP), (iii) Person and Family Committee (PFC), and (iv) national stakeholder feedback from field testing.

The Clinical Subcommittee was convened to provide detailed input on each component of the measure, and was composed of 10 members with clinical experience in ophthalmologic disease management, affiliated with 11 specialty societies, including the American Academy of Ophthalmology, American Optometric Association, American Society of Cataract and Refractive Surgery, and American Society of Retina Specialists. The Clinical Subcommittee provided recommendations on detailed specifications for the measure through an in-person meeting and a series of webinars from May 2017 to January 2018. Input was gathered in a structured manner including the use of a polling process requiring greater than 60 percent consensus.

The TEP provided high-level guidance and input on the overall direction of measure development and the framework for episode-based cost measures, while the PFC provided input on concepts of healthcare quality and value. In addition, the national field testing feedback period in October and November 2017 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, over 65,000 field test reports for TINs and TIN-NPIs were available for download and review.

---

[6] 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

One of the key roles of the Clinical Subcommittee 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 routine cataract removal, thus preventing inclusion of unrelated cost variation in this measure. Assigned services occurring in the outpatient and clinician service setting were defined separately for the pre- and post-trigger windows, and include cataract surgery, evaluation, testing, treatment, complications, and follow-up. The assigned services can also be considered as falling within five major Clinical Themes, representing categories of costs associated with the episode: cataract surgery-related office visits, complications/return to operating room, office-based diagnostic testing, office-based procedures, and other ancillary care, including anesthesia, medications, and injections.

Empirical Validity Testing

We evaluated the empirical validity of the Cataract measure by examining correlation with known indicators of resource or service utilization, specifically complications related to cataract removal. For this analysis, we compared the ratio of observed over expected spending for Routine Cataract Removal with IOL Implantation episodes with and without complications related to cataract procedures occurring in the post-trigger period. This analysis sought to confirm the expectation that variation in service utilization is captured by the Cataract measure.

In addition to the empirical validity testing above, the Scientific Methods Panel discussed the question of how different types of cost impact risk-adjusted measure scores. We address this question with further analysis into measure costs.

Certain services or costs included in the Cataract measure were classified into clinically coherent groups of services, called "clinical themes". The Cataract measure clinical themes are:

- Cataract Surgery-Related Office Visits
    - This includes E&M and eye code encounters with primary diagnosis of cataract
- Complications/Return to Operating Room
    - This includes procedures after the cataract surgery that are usually performed in an OR setting, such as IOL repositioning or exchange, removal of retained lens, IOL removal, secondary IOL placement, endothelial corneal transplant, vitrectomy, retinal detachment repair, corneal relaxing incision for surgically-induced astigmatism, amniotic membrane graft, revision/repair of surgical incision, iris/iridodialysis repair, and vitrectomy
- Office-Based Diagnostic Testing
    - This includes office-based testing prior to cataract surgery such as ultrasound, corneal pachymetry, OCT, corneal topography, ophthalmic biometry, fundus and external eye photography
- Office-Based Procedures
    - This includes procedures after the cataract surgery that are usually performed in a clinic setting, such as punctal plug placement, pneumatic retinopexy, laser barricade of retinal break, anterior chamber paracentesis, and YAG capsulotomy
- Other Ancillary Care (anesthesia, medication, injections)
    - This includes other separately-billed services performed at the time of cataract surgery or after cataract surgery, such as retrobulbar injection, anesthesia services, or intraocular injection of medications such as phenylephrine and ketorolac combination agent, vancomycin, ceftazidime, amikacin, or voriconazole

As with the original empirical validity testing, 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 took the Pearson correlation between the cost of each clinical theme and the overall risk-adjusted cost for an episode.

We expected that the Complications 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[7]. We would expect similar trends for the Office-Based Procedures theme as it contains services relating to complications, such as pneumatic retinopexy. By contrast, we expected that Office-Based Diagnostic Testing, as well as Cataract Surgery-Related Office Visits, have more nuanced, offsetting effects. While higher costs for these types of visits can directly increase the costs of an episode, research indicates that pre- and post-surgical interventions such as counselling can be associated with lower total resource use by saving on later costs[8]. Therefore, it is possible the correlation of the measure with these types of costs is lower than for complications.

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

The mean observed to expected cost for all episodes is 1.00. The mean observed to expected cost for episodes with services relating to complications during the post-trigger period is 1.04, compared with 0.95 for episodes without services relating to complications during the post-trigger period. Full results can be seen in Appendix Table 2b1.3.

Additionally, the Clinical Themes analysis demonstrates that there is a strong correlation between the Complications (correlation: 0.51) and Office-Based Procedures (correlation: 0.34) themes and risk-adjusted cost. By contrast, the Office-Based Diagnostic Testing (correlation: 0.17) and Cataract Surgery-Related Office Visits (correlation: 0.18) themes had lower correlation with risk-adjusted cost.

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

As expected, the average ratio of observed to expected cost for episodes with post-trigger complications is higher than for episodes without downstream complications. This result demonstrates that the Routine Cataract Removal with IOL Implantation measure is able to accurately capture higher resource use.

The Clinical Themes analysis demonstrates that high risk-adjusted cost is strongly associated with themes related to complications, and also linked – though more weakly, as expected -- to themes relating to testing and physician visits. 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 counselling and ophthalmologic examinations. Importantly, we see that correlation with risk-adjusted cost is strong not only for high-cost themes such as Complications (average cost for a median quintile physician: $1,166), but also for lower cost themes such as Office-Based Procedures (average cost for a median quintile physician: $345). This indicates that the correlation does not come from a mechanical increase in episode costs from high-cost themes.

2b2. EXCLUSIONS ANALYSIS

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

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

Exclusions are used in the Cataract measure to ensure a homogenous patient population within the scope of the measure focus on routine cataract removal procedures. These exclusions ensure that episodes provide meaningful information to attributed clinicians and can be seen below:

---

[7] 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. https://doi.org/10.1007/s11606-006-0254-1

[8] Devine, Elizabeth C., Cook, Thomas D., "Clinical and cost-saving effects of psychoeducational interventions with surgical patients: A meta-analysis" https://doi.org/10.1002/nur.4770090204

- Episodes where beneficiary death date occurred before the episode end.
- Episodes where the trigger claim is not performed in an outpatient hospital or ASC setting.
- Episodes where the beneficiary has ocular comorbidities.
- Episodes classified as outlier cases.

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

Given the rationale for the exclusions listed above, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a higher 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 ratio of observed over expected spending (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.

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Table 1 below presents observed cost statistics and observed over expected cost ratios for the Cataract measure exclusions. Cost statistics are also provided for the set of final episodes included in the Cataract measure for comparison, with a 10 episode case minimum at the TIN and TIN-NPI levels. Full results can be seen in Appendix Table 2b2.2.

Table 1: Cost Statistics for Routine Cataract Removal with IOL Implantation Measure Exclusions

Exclusion	Observed Cost			O/E		
	Mean	10th Percentile	90th Percentile	Mean	10th Percentile	90th Percentile
Death in Episode	$2,798	$1,751	$4,243	0.98	0.82	1.14
Episodes not occurring in HOPD or ASC Sub-Groups	$3,538	$1,923	$5,458	1.00	1.00	1.00
Ocular Comorbidity (impacting visual outcome of surgery)	$2,949	$1,764	$4,723	1.01	0.76	1.21
Ocular Comorbidity (impacting surgical complication rate)	$2,894	$1,648	$4,700	0.99	0.70	1.21
Outlier Cases	$3,614	$798	$6,622	1.25	0.33	2.39
Final Episodes (TIN)	$3,031	$1,825	$4,705	1.01	0.87	1.16
Final Episodes (TIN-NPI)	$3,033	$1,825	$4,707	1.01	0.87	1.16

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Although statistical results indicate that the excluded episodes, aside from outliers, have fairly similar results to the final set of episodes, these episodes were still excluded due to clinical considerations to ensure a comparable patient cohort that will yield meaningful information to attributed clinicians. Further discussion of the results and applicable clinical rationale for each exclusion is outlined below.

Episodes ending in death: Cases where the beneficiary dies during the episode are not eligible to be included in the Cataract measure. Though the difference between mean observed episode cost for death and non-death episodes is relatively small ($2,798 compared to $3,031), there are a few explanations for the exclusion of death episodes. Including episodes ending in death in the measure calculation may distort measure scores. The observed cost for episodes ending in death is less than the expected cost because these are likely to be shorter episodes (and therefore include fewer services) than beneficiaries with episodes that do not end in death. Because of this, including episodes ending in death in measure calculation may distort measure scores where truncated periods of care give the appearance of more cost effective care. Related to this, the measure seeks to avoid problematic incentives that could arise with the inclusion of episodes ending in death that lower a measure score.

Episodes where the trigger claim was not performed in an HOPD or ASC setting: Cases where the trigger claim occurred in an inpatient hospital or office setting are not included in the Cataract measure. The mean observed cost of these episodes is approximately $500 more than for the final set of episodes. This is in line with expectations and expert clinical input about the different patient cohort that has this procedure in settings outside of HOPD and ASC. As such, these cases are not included in the measure to ensure a clinically comparable patient cohort.

Episodes where the beneficiary has ocular comorbidities (impacting visual outcome of surgery or surgical complication rate: While the mean observed cost is slightly lower for these episodes than the final set of episodes included in the measure, the presence of ocular comorbidities in a patient is outside the scope of the measure intent to capture uncomplicated, routine cataract removal procedures. In addition, the use of the ocular condition exclusions based on two NQF-endorsed quality measures for outcomes of cataract surgery or complications rate is intended to allow attributed clinicians to receive feedback on the same patient cohort for quality and cost measures.

Outlier cases: Outliers are excluded from the Cataract measure calculation to avoid cases where a handful of high-cost and low-cost outliers have a disproportionate effect on measure score. The ratio of observed to expected episode cost ranges from 0.33 at the 10th percentile to 2.39 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 the spending levels one would have expected based on patient characteristics.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4**.**

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 99 risk factors**

☒ **Stratification by 4 risk categories**

☐ **Other,**

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

The Routine Cataract Removal with IOL Implantation measure risk adjustment model 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 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 measure also includes variables for factors that affect resource use that expert clinician input identified as important to account for.

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. The 79 HCC indicators are specified in the CMS-HCC Version 22 (V22) 2016 model. Patients without a full 120-day lookback period (i.e., the beneficiary is not enrolled in both Medicare Part A and Medicare Part B for the 120 days prior to the episode trigger) have their episodes excluded from the measure. This 120-day period is used to measure 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.

The risk adjustment methodology includes status indicator variables for whether the beneficiary qualifies for Medicare through Disability or End-Stage Renal Disease (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 measures of severity of illness indicators.

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

The Routine Cataract Removal with IOL Implantation measure risk adjustment model also includes additional factors to further isolate costs that attributed clinicians can reasonably influence, informed by recommendations from the Clinical Subcommittee based on their clinical expertise and empirical analysis. These additional risk adjustors capture:

(i) whether the beneficiary is a new or established patient

(ii) whether the procedure was performed in part by a resident under the direction of a teaching physician

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

The risk adjustment model outlined above is performed separately for each of the four Cataract measure sub-groups which are based on the procedure place of service (i.e., ambulatory surgery center [ASC] or hospital outpatient department [HOPD]) and laterality:

(i) ASC – Unilateral

(ii) ASC – Bilateral

(iii) HOPD – Unilateral

(iv) HOPD -- Bilateral

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

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

N/A

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

The CMS-HCC model was selected 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 (FFS) 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 on adapting the CMS-HCC model to the Cataract measure methodology.

Measure-specific risk adjustors were selected based on expert clinician input through the Clinical Subcommittee. Members were provided with empirical analyses on different 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 of clinician control, or any other factors that would help prevent unintended consequences. For this measure, Subcommittee members recommended accounting for the following variables to avoid unintended consequences:

(i) whether the beneficiary is a new or established patient, accounting for referral-based practices that may disproportionately treat new patients, and;

(ii) whether the procedure was performed in part by a resident under the direction of a teaching physician, accounting for cost differences from clinicians engaged in teaching resident surgeons.

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 selected based on expert recommendation from the Clinical Subcommittee, with the goal of ensuring clinical comparability among episodes so that the cost measure fairly compares clinicians with similar patient case-mix. The Clinical Subcommittee recommended the following Routine Cataract Removal with IOL Implantation sub-groups:

(i) ASC – Unilateral

(ii) ASC – Bilateral

(iii) HOPD – Unilateral

(iv) HOPD -- Bilateral

The stratification for site of service accounts for access factors, as some clinicians may not have access to an ASC, a lower-cost setting than HOPD, due to regional availability or as a result of health plan contracting arrangements. Sub-groups for unilateral and bilateral surgery are used to account for scenarios where some services may be applied to a second surgery performed in close succession, meaning that bilateral procedures will likely be more expensive than unilateral ones. More information on sub-groups can be found in Section 2b3.9 on risk stratification analyses.

Information on data sources and methodologies used to analyze social risk factors can be seen in Section 1.8 of this testing form.

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

☒ **Published literature**

☒ **Internal data analysis**

☐ **Other (please describe)**

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

The Routine Cataract Removal with IOL Implantation risk adjustment model broadly replicates the CMS-HCC V22 2016 model. 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, more information on factors included in the CMS-HCC model can be found at Pope et al. 2011.[9] Expert clinician input was also sought through a Clinical Subcommittee, including recommendations on additional risk adjustors and sub-groups.

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

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

Acumen analyzed gender, dual status, income, education, and unemployment as social risk factors (more information on these variables can be found in Section 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 61 percent to 65 percent across the four sub-groups in this measure. The majority of the beneficiaries (84% - 92%) 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. While two to three percent of beneficiaries are classified below a high school education level, the majority of episodes are classified at a high school level or greater. Finally, 20 to 23 percent of beneficiaries have high unemployment designation (>10%). Full results can be seen in Appendix Table 2b3. 4b.1.

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 adjustors described in previous sections (variables from the CMS-HCC V22 2016 model, disability status, ESRD status, interaction variables, recent long-term care use, and measure-specific clinical risk adjustors). 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

[9] 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.

revealed many significant p-values, indicating that social risk factors are likely predictive factors for determining resource use among beneficiaries for the relevant characteristic. However, the analysis also shows that the directions of the effects of social risk factors are not consistent. For example, high income episodes may display higher spending for the ASC/Unilateral sub-group but lower spending for the other three sub-groups. Full results can be seen in Appendix Table 2b3.4b.1 and Table 2b3.4b.2.

Secondly, we analyzed the impact of adding these social risk variables on overall model performance by looking at the differences in the ratio of observed to expected episode cost (O/E) 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 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 measure scores for 99 percent of TINs and 99 percent of TIN-NPIs changed by ±0.03 or less. Please see Appendix Table 2b3. 4b.3 for complete results.

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 level, with a Pearson correlation coefficient of 0.999, and the TIN-NPI level with a correlation coefficient of 0.999. 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 Cataract measure risk adjustment model sufficiently accounts for the effects of social risk factor on clinician measure scores.

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

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

If stratified, skip to 2b3.9

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

(a) *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 discussed in Section 2b3.6.

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

(c) *Coefficient estimates, standard errors, and p-values* were 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.

Predictive ratios are provided to aid in the overall assessment of the predictive ability of the risk adjustment models (Table 2b3.8). The coefficient results are provided in Appendix Table 2b3.1.1 and Table 2b3.8.

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

The overall R-squared for the Routine Cataract Removal with IOL Implantation cost measure , calculated by dividing explained sum of squares by total sum of squares is 0.75. The adjusted R-squared is also 0.75.

Appendix Table 2b3.1.1 also includes regression coefficients and standard errors for each of the covariates used in the risk adjustment models. More information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.[10]

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

We interpret calibration as how accurately the risk model's predictions match the actual episode cost. For each of the risk factors included in the model, we calculate the average observed cost over expected cost ratio to demonstrate the model's prediction accuracy. The average observed to expected cost is generally close to one across risk factors, indicating that the model is accurately predicting actual episode cost for that risk factor. Full results can be seen in Appendix Table 2b3.1.1.

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

As seen in Appendix Table 2b.3.8 showing predictive ratios by risk decile for the measure, the model has consistent predictive ratios across risk score deciles, with each decile having a predictive ratio of 1.00.

2b3.9. Results of Risk Stratification Analysis:

Results indicate that the four measure sub-groups have varying measure scores (see Appendix Table 2b4.2). Specifically, bilateral cases are more expensive than unilateral cases. At the TIN level, the mean score for bilateral HOPD episodes is $5,066 compared to unilateral HOPD episodes at $2,696. This trend is also seen for bilateral ASC episodes with a mean score of $3,882 compared to $2,092 for unilateral ASC episodes. Results are similar at the TIN-NPI level. Thus, bilateral cases are considered separate from unilateral cases (and are not considered as two unilateral cases given that services such as diagnostic or pre-operative care cannot be easily divided).

HOPD cases are more expensive than ASC cases. At the TIN level, the mean measure score for HOPD unilateral cases is $2,696 compared to $2,092 for ASC unilateral cases. Similarly, HOPD bilateral episodes have a mean measure score of $5,066 compared to ASC bilateral episodes with a mean score of $3,882. The trend is the same at the TIN-NPI level. Given that clinicians often do not have access to the same treatment options, ASC and HOPD episodes are considered separately to prevent bias against clinicians who may not have access to ASC. Whether an episode occurs in an ASC or HOPD may also be related to the beneficiaries' underlying characteristics such as health conditions. Stratifying episodes into these sub-groups helps ensure meaningful comparison of clinician resource use.

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

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.[11] As noted in Section 2b3.5, 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.

[10] Ibid.

[11] Ibid, 6.

As demonstrated in Section 2b3.7 and 2b3.8, the average observed over expected ratios for each risk factor included in the model and 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 individual risk factors or overall risk level.

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

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

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

Our method to determine 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.

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

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

(i) the 99th percentile of the measure score is nearly two times the 1st percentile at both the TIN level and TIN-NPI levels;
(ii) the Cataract measure score at the 90th percentile is approximately 20 percent greater than the score at the 10th percentile at both the TIN and TIN-NPI level; and
(iii) the mean Cataract score for episodes in HOPD sub-groups is approximately 30 percent higher than for episodes in ASC sub-groups at both the TIN and TIN-NPI levels.

These results indicate there is large 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 $100 range (i.e., $2,995-$3,078 at the TIN level and $2,981-$3,062 at the TIN-NPI level). Similarly, clinicians in urban areas seem to perform comparably to those in rural areas.

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 median TIN score of $2,938 to $3,107 and a range in median TIN-NPI score of $2,928 to $3,112, indicating that the risk adjustment model is overall functioning as intended. Full results can be seen in Appendix Table 2b4.2.

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

There are clinically and practically significant variation in Cataract measure scores, indicating the measure's ability to capture differences in performance. The measure was constructed with extensive, detailed clinical input to assign only services that are related to the procedure; as such, the differences in costs across clinicians are limited to costs that are within the reasonable influence of the attributed clinician. This leads to a more clinically meaningful and actionable comparison of cost across TINs and TIN-NPIs.

Our findings regarding variation in measure scores are consistent with expert clinician input. The Ophthalmologic Clinical Subcommittee suggested development of sub-groups based on place of service, noting the differences in cost between ASCs and HOPDs and the ease of access to ASCs is dependent on regional availability or other factors. The results show minimal (only a two percent) difference in cost between rural and urban location since those differences are already accounted for through the clinically meaningful sub-groups that considers regional availability of these places of services.

Overall, results expectedly show that clinicians are not being systematically penalized or rewarded due to risk score decile given the current Cataract measure design.

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

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

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

Since CMS uses Medicare claims data to calculate the Cataract 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 Enrollment Database (EDB), the beneficiary does not appear in the EDB, or the beneficiary death date occurs before the episode trigger date.

The Cataract 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 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.

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

The table below presents the frequency of missing data across the four categories of missing data which caused episodes to be excluded from the Cataract 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 table also shows these figures as percentages of all episodes, TINs, or TIN-NPIs. 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 2: Missing Data Categories for the Routine Cataract Removal with IOL Implantation Measure

Exclusion	# Episodes	# TINs	# TIN-NPIs
Missing birth date	0	0	0
Death before trigger	*	*	*
Other primary payer	175,515	5,122	9,837
Not continuously enrolled	105,510	4,921	9,338

Please note that values with an asterisk in the table above indicate there were fewer than 11 episodes or beneficiaries in that category.

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

Since the Cataract 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 ensure that missing or inaccurate information in claims data is not included in the cost measure.

Feasibility

F.1. Byproduct of Care Processes

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

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

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

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

If other:

F.2. Electronic Sources

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

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

ALL data elements are in defined fields in electronic claims

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

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

Attachment:

F.3. Data Collection Strategy

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

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

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.

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. Therefore, the time at which a measure developer pulls claims data represents a trade-off between efficiency (accessing the data soon) and accuracy (waiting until most claims are finalized). In order to determine the appropriate "run-out" period for claims data, 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 purposes. MIPS reporting for this cost measure will be done in line with program reporting.

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 (see Section 2b6 of the measure testing form for addition details). 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.

Sampling

During measure testing, Acumen noted that episodes in which the beneficiary died prior to the episode end date exhibited different cost distributions to other episodes. In order to avoid this effect impacting 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.

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

N/A.

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

Usability and Use

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

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

U.1.1. Current and Planned Use

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

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

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

Program Name: Quality Payment Program (QPP) Merit-based Incentive Payment System (MIPS)
Sponsor: CMS
Purpose: The Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) established the Quality Payment Program. Under the Quality Payment Program, clinicians are incentivized to provide high-quality and high value care through Advanced Alternate Payment Models (APMs) or the Merit-based Incentive Payment System (MIPS). MIPS eligible clinicians will receive a performance-based payment adjustment to their Medicare payment. This payment adjustment is based on a MIPS final score that assesses evidence-based and practice-specific data across the following categories:

1. Quality
2. Improvement activities
3. Advancing care information
4. Cost

As specified in the CY 2019 Physician Fee Schedule final rule (83 FR 59765 through 59776), this measure will be implemented as part of MIPS beginning in the 2019 MIPS performance year and 2021 MIPS payment year.
Geographic Area: U.S.

Number/Percentage of Accountable Entities: The number of clinicians in the Quality Payment Program varies by performance period. For 2017, there were 1,057,824 MIPS eligible clinicians receiving a MIPS payment adjustment.[1] As clinicians have choices on how to participate in the Quality Payment Program (e.g., through MIPS or the Advanced APMs, as groups or individuals), the exact number and percentage of clinicians who will receive a performance score on this measure will only be confirmed after the end of each performance period. The number of patients covered by this measure is dependent on whether providers report at the group (TIN) or individual clinician (TIN-NPI) level. The number of patients covered by group reporting is 490,717, while the number of patients covered by individual reporting is 485,228.

[1] CMS, 2017 Quality Payment Program Reporting Experience, https://qpp-cm-prod-content.s3.amazonaws.com/uploads/491/2017%20QPP%20Experience%20Report.pdf

Number/Percentage of Accountable Patients: N/A.

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

N/A.

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

N/A.

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

DEVELOPMENT: FIELD TESTING

Acumen and CMS conducted a national field test of episode-based cost measures, including the Cataract measure, for a 35-day comment period (October 16 to November 20, 2017). The testing sample for providing field test reports was all clinicians and clinician groups who were attributed 10 or more episodes associated with at least one of eight episode-based cost measures developed during 2017. The measurement period was June 1, 2016, to May 31, 2017. Cost performance information on these episodes was provided in confidential reports available for download on the CMS Enterprise (EIDM) Portal by attributed clinicians and clinician groups.[1]

A sample of 17,557 clinician groups and 48,263 clinicians received a confidential field test report. The testing sample was selected to balance coverage (i.e., a lower minimum number of episodes per attributed clinician increases the number of TIN-NPIs and TINs who would receive reports) and reliability (i.e., a higher minimum number of episodes per clinician or clinician group provides more reliable and meaningful metrics), 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 did not determine case minimums used for program implementation.

We provided field test reports for the following number of clinician groups and clinicians. Each report included information for all measures for which the clinician or clinician group was attributed 10 or more episodes.

- Total: 17,557 TINs; 48,263 TIN-NPIs
- Routine Cataract Removal with IOL Implantation: 4,434 TINs; 7,690 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 Frequently Asked Questions document, and a Fact Sheet.[2] During field

testing, Acumen conducted education and outreach activities including National Provider Call webinars, office hours with specialty societies, and Help Desk support.

The purpose of field testing was to provide a voluntary opportunity for clinicians and other stakeholders to provide feedback on: (i) the draft measure specifications, including each component of the measure (e.g., the clinical validity of assigned services and the trigger codes), (ii) the field test report template (e.g., what information is most meaningful to allow clinicians to make changes to their care practices), and (iii) all accompanying documentation (e.g., the level of detail in specifications documentation). Acumen sought feedback through an online survey, with the option to attach a comment letter in PDF or Word document format.

[1] CMS Enterprise Portal, https://portal.cms.gov/wps/portal/unauthportal/home/

[2] These documents were posted to the CMS MACRA Feedback page (https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/MACRA-Feedback.html). The field testing fact sheet and FAQs are in a zip file at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2017-field-test-materials.zip.

IMPLEMENTATION: PRE-RULEMAKING and RULEMAKING

The Cataract measure was implemented in MIPS after going through the pre-rulemaking process and notice-and-comment rulemaking. The measure was submitted to and included in the 2017 Measures Under Consideration (MUC) List. It was then considered by National Quality Forum (NQF)'s Measure Applications Partnership (MAP) Clinician Workgroup and Coordinating Committee in December 2017 and January 2018, respectively. The final recommendation from the MAP was 'conditional support for rulemaking,' with the condition of NQF endorsement.

The measure was proposed for use in the MIPS cost performance category in the CY 2019 Physician Fee Schedule proposed rule.[1] Measure specifications were publicly posted and linked to from the proposed rule. A National Summary Data Report containing information about the measure performance (e.g., measure score distributions by different provider characteristics) was also publicly posted. Stakeholders submitted comments on the proposed rule during a 60-day public comment period. CMS considered these comments and finalized the measure for use in MIPS from the CY 2019 performance period onwards in the CY 2019 Physician Fee Schedule final rule.[2]

[1] The CY 2019 Physician Fee Schedule proposed rule can be found here: https://www.federalregister.gov/documents/2018/07/27/2018-14985/medicare-program-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other-revisions

[2] The CY 2019 Physician Fee Schedule final rule can be found here: https://www.federalregister.gov/documents/2018/11/23/2018-24170/medicare-program-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other-revisions

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

FIELD TESTING:

National field testing was organized for the purpose of gathering targeted comments on the Cataract measure. During the feedback period, field test reports were accessed by accounts corresponding to a total of 1,364 clinician groups (TINs) and 10,628 clinicians (TIN-NPIs). After field testing, the comments received on the measure were summarized for the Clinical Subcommittee to consider in making refinements. Field test reports continued to be available on the CMS Enterprise Portal until September 2018.

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 Excel file contained the following sheets, which were described in more detail in Appendix C of the field test reports ("How to Interpret this Report"):

- Summary
    - High-level information on the performance of the TIN or TIN-NPI across all episodes within each measure attributed to said TIN or TIN-NPI
    - Metrics listed in this tab related to the cost measure score:
        - Episode count
        - Average episode risk score percentile
        - Cost measure score (average risk-adjusted cost to Medicare for that measure)
        - National average cost measure score and percent difference between the TIN/TIN-NPI's score and the national average
- Results for Each Measure
    - Understanding Your Cost Measure Score: information from the Summary tab in context
    - Breakdown of Cost Measure Score by Episode Sub-Group: comparison of the TIN/TIN-NPI´s average risk-adjusted cost to Medicare to the national average risk-adjusted cost to Medicare for the measure as a whole and separately for each episode sub-group
        - Episode sub-groups are divisions within a cost measure's episode group that define more homogenous patient cohorts to ensure clinical comparability (i.e., the cost measure fairly compares like patients)
    - Breakdown of Episodes by Episode Sub-Group for Your TIN/TIN-NPI and National Average: comparison of the allocation of the TIN/TIN-NPI's episodes to the various sub-groups within the overall episode group to the average allocation across episodes for TINs/TIN-NPIs nationally
    - Breakdown of Part B Physician/Supplier Episode Cost by Your TIN/TIN-NPI vs. Other TINs/TIN-NPIs: average share of episode costs that came from the evaluated TIN/TIN-NPI versus other TINs/TIN-NPIs and average of each share across episodes for TINs/TIN-NPIs nationally
    - Breakdown of Utilization and Cost by Selected Clinical Theme: TIN´s/TIN-NPI's service utilization and costs by "clinical themes" (clinical categorizations of the services assigned to episode costs during the episode window)[1]
- Appendix A for Each Measure
    - More detailed information on potential cost drivers in the TIN/TIN-NPI's episodes
    - Breakdown of Utilization and Cost by Medicare Setting and Service Category: analysis of utilization and cost for the measure, both for all services and by specific service categories[2]
    - Breakdown of Utilization and Cost for Physician/Supplier Part B Claims: same comparison of utilization and cost as given in "Breakdown of Utilization and Cost by Medicare Setting and Service Category" above (i.e., (i) the national average, (ii) TINs/TIN-NPIs in the same risk bracket, and (iii) the evaluated TIN/TIN-NPI), but by top 5 most billed services and by risk bracket
    - Breakdown of Utilization and Cost for Inpatient Claims: same information as in "Breakdown of Utilization and Cost for Physician/Supplier Part B Claims" for inpatient claims assigned to the TIN/TIN-NPI's episode costs
- Appendix B
    - Detailed episode-level 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 Within the Episode

[1] Definitions of the clinical themes are available in the "SA_" tabs of the Measure Codes List file for the measure, downloadable from the QPP Resource Library at this link: https://qpp-cm-prod-content.s3.amazonaws.com/uploads/344/2019%20Cost%20Measure%20Code%20Lists.zip

[2] Definitions of the various categories of services presented in this table can be found on page 438, Table C.2 of the "Detailed Methods of the 2015 Supplemental Quality and Resource Use Reports (QRURs)" document available here: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Downloads/2015-SQRUR-Detailed-Methods.pdf

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 Clinical Subcommittee recruitment efforts, as well as CMS, QPP, and other available listservs. Outreach emails included:

- Targeted messages to a small number of specialty societies whose members we anticipated would be attributed a report, to assist in reaching their members about field testing
- Targeted emails to available contact details linked to a TIN or TIN-NPI that received a field test report
- General emails to contacts from clinician and healthcare provider organizations, noting that we sought feedback from all stakeholders even if they did not receive a confidential report
- General emails to all our contacts in clinician and healthcare provider organizations to inform them about the opportunity to join National Provider Calls (NPCs)

Acumen and CMS hosted two office hours sessions on September 14 and 18, 2017, to provide an overview of field testing to specialty societies, discuss what information their members would be particularly interested in, and answer any questions. After the webinars, Acumen also prepared and distributed measure summaries that societies could use to inform their members of the basic specifications on which Acumen was requesting input. Between the two webinars, there were 31 attendees affiliated with at least 21 specialty societies, including representatives from the American Society of Cataract and Refractive Surgery, the American Academy of Ophthalmology, and the American Medical Association, among others.

During the field testing feedback period, Acumen organized an inquiry management strategy with the Physician Value and QPP Help Desks; Acumen directly handled more than 160 inquiries during the feedback period.

Acumen and CMS hosted two National Provider Calls on October 30, 2017, and November 2, 2017, to engage clinicians and other stakeholders during field testing. The two webinars, both covering the same content, consisted of an hour-long presentation, outlining (i) the cost measure development activities, (ii) how to access the confidential field test reports, and (iii) the contents of the reports. The presentation was followed by a 30-minute Q&A session. In total, approximately 1,000 people attended one of the webinars and around 120 comments and questions were received via webinar chat and on the phone.

PRE-RULEMAKING:

There was a public comment period after the release of the Measures Under Consideration (MUC) list from November 30, 2017 to December 7, 2017, prior to the MAP Clinician Workgroup meeting. The MAP Clinician Workgroup met on December 12, 2017, to consider measure specifications and testing updates. In accordance with MAP procedure, these documents were not publicly released but were made available to MAP members. Following the release of the Clinician Workgroup's preliminary recommendation, the report was open for a public comment period from December 21, 2017, to January 11, 2018. The MAP Coordinating Committee met on January 25-26, 2018, to consider these comments alongside the Clinician Workgroup's recommendation. Both MAP meetings were open to the public.

RULEMAKING:

During the public comment period for the proposed rule from July 12, 2018, to September 10, 2018, stakeholders could review the proposed rule language, measure specifications, and National Summary Data Report when submitting comments. CMS conducted email outreach via its listserv to notify stakeholders about the release of the proposed rule.

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

FIELD TESTING:

In total, Acumen received 219 survey responses and 53 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 proved advantageous in reaching a wider audience, increasing the amount and variety of feedback provided, and facilitating a faster turnaround for the measure development team to process and operationalize feedback. The survey was divided into four sections for general and detailed questions on the reports themselves, questions on the supplemental documentation, and questions on the measure specifications. Questions in the survey included Likert scales specific to the report, process, and measure components; multiple-choice questions; and open response questions. The survey was designed to take 20-30 minutes, but allowed flexibility based on a stakeholder's use of open-ended responses and the number of measures on which they chose to provide feedback.

Inquiries were also registered and feedback submitted via email to macra-episode-based-cost-measures-info@acumenllc.com, Physician Value and QPP Help Desks, and verbally and via webinar chat at the NPCs and office hours.

PRE-RULEMAKING:

CMS received over 40 comments on the eight episode-based cost measures included in the 2017 Measures Under Consideration List. This included seven comments for the Routine Cataract Removal with Intraocular Lens (IOL) Implantation Cost Measure. After the MAP Clinician Workgroup meeting in December 2017, there was another public comment period on their preliminary recommendations, which received over 20 comments across the eight measures, with three comments specific to the Routine Cataract Removal with Intraocular Lens (IOL) Implantation Cost Measure. These public comment periods were facilitated by NQF. Stakeholders were able to submit their comments via the NQF website.

RULEMAKING:

CMS received over 15,368 comments on the CY 2019 Physician Fee Schedule proposed rule. A search on the regulations.gov website returns 242 results for "episode-based cost measure" as a rough approximation of the number of comments on the eight episode-based cost measures during rulemaking. Stakeholders could submit comments through the Federal Register website or via mail.

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

FIELD TESTING:

The publicly available Field Testing Feedback Summary Report[1] 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 is an important aspect of the measure development process. Stakeholders expressed appreciation for the opportunity to provide feedback during field testing and for CMS' continued effort to involve stakeholders in the measure development process, such as convening Clinical Subcommittees to seek an extensive amount of clinical input in constructing these measures. Commenters urged CMS to continue to work closely with specialty societies and other involved stakeholders.

- Provide additional time for stakeholders to review materials and provide feedback during field testing. According to some stakeholders, the October to November 2017 field testing feedback period was too short given the large amount of new information that was presented and suggested that the period be extended or be kept open.
- Accessing the confidential field test reports from the CMS Enterprise Portal presented many challenges. Some stakeholders noted that they faced difficulties creating accounts and downloading their confidential field test reports from the portal that may have had a negative impact on the number of clinicians who took part in field testing.
- While some stakeholders believed the field test report presented useful information for understanding clinician cost measure performance, they also highlighted areas for improvement in regard to providing actionable information. Stakeholders praised the navigability and the inclusion of useful information in the report. However, some stakeholders also expressed concerns with the comprehensibility of the report and its usefulness in terms of providing actionable information for clinicians.
- Stakeholder feedback received on the supplemental field testing materials was mixed, with some stakeholders finding them helpful and informative and others believing the materials were too complex. Some stakeholders found the supplemental field testing materials informative, providing helpful information on field testing and the specifications of the cost measures. Some stakeholders believed that the materials were not detailed enough. However, many noted that the materials were comprehensive but too lengthy and complex, and they believed the amount of information was overwhelming to absorb within the field testing feedback period.

The aforementioned report additionally contains measure-specific feedback, which was used as the basis for the post-field testing measure refinements discussed in U.2.3. At a high level, feedback included the following recommendations:

- 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.
- Feedback collected via email, Help Desk, and at the NPCs and office hours covered a wide range of topics, such as:
- Email and Help Desk inquiries: accessing field test reports, MIPS cost performance category, cost measures for chronic conditions, interpreting field test reports, using the online survey, payment standardization
- NPC and office hours comments and questions: risk adjustment methodology, supplemental field test resources, field test methodology, quality alignment, future cost measures

[1] The report can be downloaded at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2018-field-testing-feedback-summary-report.pdf

PRE-RULEMAKING:

The MAP gives feedback on performance measures from a wide variety of perspectives, with representatives including “consumers, businesses and purchasers, laborers, health plans, clinicians and providers, communities and states, and suppliers.”[2] The Clinician Workgroup specifically aims to “ensur[e] the alignment of measures and data sources to reduce duplication and burden, identif[y] the characteristics of an ideal measure set to promote common goals across programs, and implemen[t] standardized data elements.”[3]

[2] National Quality Forum, Measure Applications Partnership
https://www.qualityforum.org/Setting_Priorities/Partnership/Measure_Applications_Partnership.aspx

[3] National Quality Forum, MAP Member Guidebook
http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80515

RULEMAKING/PUBLIC COMMENT:

CMS received comments on the proposed episode-based cost measures during the public comment period for the CY 2019 Physician Fee Schedule proposed rule. There was support from several commenters on the proposed adoption of the episode-based cost measures in MIPS. Commenters provided feedback on the development process, including voicing support for the development of episode-based cost measures through a transparent process that engages with stakeholders and submitting critiques of the short timeline clinicians are given to understand and gain experience with the measures before they are used in the program. CMS also received comments supporting the submission of the episode-based cost measures for NQF endorsement prior to their use in the program. Measure-specific comments were also received on the specifications of the measures, which CMS and Acumen reviewed to determine whether changes needed to be made to the specifications of the measures. For more detailed information on the comments received on the measures as part of the proposed rule public comment period, please see the episode-based cost measures section in the CY 2019 Physician Fee Schedule final rule for a summary of the public comments received along with CMS responses: https://www.federalregister.gov/d/2018-24170/p-2965.

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

PRE-RULEMAKING:

The MAP recognized the importance of cost measures to the MIPS program and conditionally supported the Routine Cataract Removal with Intraocular Lens (IOL) Implantation cost measure pending NQF endorsement. During the NQF endorsement review, the MAP encouraged the Cost and Resource Use Standing Committee to specifically 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 Standing Committee should also examine the exclusions in this measure to ensure appropriate attribution.

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

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 a Clinical Subcommittee comprised of subject matter and measure-development experts.

After completing field testing, the feedback provided through the survey and comment letters was compiled into a measure-specific report, which was then provided to the Clinical Subcommittee (CS) that provided the bulk of measure development input. CS members then discussed and voted on which of the proposed specifications updates should be implemented in the finalized measure. More specifically, this process included the following steps for each of two webinars:

- Pre-webinar production of summary sheets, first of applicable field testing feedback and then of first-round measure refinements
- The webinar itself, to gather first substantive and then non-substantive measure refinement feedback
    - CS members discussed each update suggested by field testing commenters to determine whether, based on their best clinical input, they should recommend implementation of the change or not.

    - In some cases, CS members acknowledged the validity of the suggestion, but felt they had already addressed the commenter(s) concerns.
- A survey to gather CS member input
    - For the purposes of considering which measure specifications changes to implement, CS consensus was defined as >60% agreement.
- Incorporation of CS input into final measure specifications
- The changes to the Cataract measure made as a result of field testing feedback are as follows:
- Sub-Grouping: Update sub-groups to remove co-management group, resulting in the following sub-groups:

(1) ASC/Bilateral

(2) ASC/Unilateral

(3) HOPD/Bilateral

(4) HOPD/Unilateral

- Service Assignment: Only include pre-operative visits which are billed with a cataract diagnosis (primary or secondary)
- Risk Adjustment: Add risk adjustors for:
    - Episodes billed with a GC modifier code
    - Episodes with new patient E&M codes versus episodes with only established patient E&M codes
- Exclusions: Eliminate exclusion of (i.e., no longer exclude) H26.9 Unspecified Cataract

RULEMAKING/PUBLIC COMMENT:

During the public comment period for the CY 2019 Physician Fee Schedule proposed rule, stakeholders submitted comments on the proposed episode-based cost measures, including on the Cataract measure. While we received feedback on the proposed measures generally, as described in Section U.2.2.2, there was no measure-specific feedback received on the specifications of this measure. Therefore, the measure was finalized as proposed.

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

Discuss:

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

N/A.

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

N/A.

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

While the measure has technically been implemented into the MIPS program, the measure results are first scheduled to be calculated for performance year 2019 (payment year 2021), and thus no unexpected consequences can be identified at this time.

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

While the measure has technically been implemented into the MIPS program, the measure results are first scheduled to be calculated for performance year 2019 (payment year 2021), and thus no unexpected consequences can be identified at this time.

Related or Competing Measures

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

H.1. Relation to Other NQF-endorsed Measures

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

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

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

Related NQF-Endorsed Measures:

There are no NQF-endorsed cost measures with the same focus or the same target population.

Competing NQF-Endorsed Measures:

There are currently no NQF-endorsed measures that address both this same measure focus AND this same target population.

Related Non-NQF-Endorsed Measures:

There are no non-NQF-endorsed cost measures with the same focus or the same target population submitted to NQF or implemented in MIPS.

Competing Non-NQF-Endorsed Measures:

There are currently no non-NQF-endorsed measures that address both this same measure focus AND this same target population.

H.2. Harmonization

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

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

H.3. Competing Measure(s)

H.3.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

N/A. There are currently no measures that address both this same measure focus AND this same target population. This Cataract measure evaluates clinicians' and clinician groups' risk-adjusted episode cost. The target population is Medicare beneficiaries enrolled in Medicare fee-for-service and who undergo a procedure for routine cataract removal with IOL implantation that triggers a Routine Cataract Removal with IOL Implantation episode. The cohort for this cost measure is also further refined by the definition of the episode group and measure-specific exclusions.

Contact Information

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

Co.2 Point of Contact: Joel, Andress, joel.andress@cms.hhs.gov, 410-786-5237-

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

Co.4 Point of Contact: Binglie, Luo, ccsq-macra-support@acumenllc.com, 650-558-8882-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

Acumen convened multiple stakeholder and expert groups to contribute to the measure development process, including Clinical Subcommittees and a Technical Expert Panel (TEP). Clinical Subcommittees convened between May 2017 and January 2018 made recommendations on all components of the episode-based cost measures, including what diagnoses and/or procedures should trigger and define an episode, which services should be assigned to an episode, what patient populations should be excluded, and which clinical characteristics should be accounted for in the risk adjustment model. The TEP, which met four times between August 2016 and August 2017, served a high-level advisory role and provided cross-measure guidance on the overall direction of measure development.

Technical Expert Panel Members:

Adolph Yates, American Academy of Orthopaedic Surgeons

Alan Lazaroff, American Geriatrics Society

Allison Madson, American Society of Cataract and Refractive Surgery

Alvia Siddiqi, American Academy of Family Physicians

Anupam Jena, Harvard Medical School

Caroll Koscheski, American College of Gastroenterology

Chandy Ellimoottil, American Urological Association

Diane Padden, American Association of Nurse Practitioners

Dyane Tower, American Podiatric Medical Association

Edison A. Machado, Jr., The American Health Quality Association

Jackson Williams, Dialysis Patient Citizens

James Naessens, Mayo Clinic

John Bulger, American Osteopathic Association

Juan Quintana, American Association of Nurse Anesthetists

Kata Kertesz, Center for Medicare Advocacy

Kathleen Blake, American Medical Association

Mary Fran Tracy, National Association of Clinical Nurse Specialists

Parag Parekh, American Society of Cataract and Refractive Surgery

Patrick Coll, University of Connecticut Health Center

Shelly Nash, Adventist Health System

Sophie Shen, Johnson and Johnson Health Care Systems, Inc.

Ophthalmologic Disease Management Clinical Subcommittee Members:

Andrew Morgenstern, American Optometric Association

April Maa, American Academy of Ophthalmology

Cynthia (Cindie) Mattox, American Academy of Ophthalmology

David Glasser, American Academy of Ophthalmology

John Hitchens, American Association of Nurse Anesthetists

John Thompson, American Society of Retina Specialists

Mark Levine, American Geriatrics Society

Parag Parekh, American Society of Cataract and Refractive Surgery

Peter Goldzweig, American Society of Anesthesiologists

Scott Friedman, American Academy of Ophthalmology

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision:

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

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

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: