Quality Payment
PROGRAM

Cataract Removal with Intraocular Lens (IOL) Implantation

Measure Testing Form

February 2023

Contents

1.0 Introduction **3**
1.1 Project Title and Overview 3
1.2 Measure Name 3
1.3 Type of Measure 3
1.4 Data 3
2.0 Preliminary Testing Results **4**
2.1 Impacts of Revisions to the Measure 4
2.1.1 Measure Coverage 4
2.1.2 Frequently Attributed Specialties 5
2.1.3 Reliability 5
2.2 Validity 6
2.3 Performance Gap 8
2.4 Risk Adjustment and Stratification 9
2.4.1 Discrimination 10
2.4.2 Calibration 10
2.5 Social Risk Factor Analysis 10
2.6 Impact of Exclusions 15
Appendix A. Distributions of Measure Score (Revised Measure) **17**
Appendix B. Associations between Social Risk Factor Variables and Cost of Care for TIN-NPIs **18**

1.0 Introduction

This Measure Testing Form (MTF) provides a brief summary of the preliminary measure testing results as part of the comprehensive reevaluation process for three episode-based cost measures. Readers may review these results, alongside other documentation, to provide feedback on the measure using the comprehensive reevaluation survey. The testing results reflect both the version of the measure that is currently in-use in MIPS and a revised version of the measure that is undergoing updates potential use in MIPS in future years. Please see the Draft Cost Measure Methodology for a description of the measure specifications and the Draft Measure Codes List for the list of codes used to specify the measure.[1]

1.1 Project Title and Overview

The Centers for Medicare & Medicaid Services (CMS) has contracted with Acumen, LLC to develop and maintain episode-based cost measures for potential use in the Merit-Based Incentive Payment System (MIPS) to meet the requirements of the Medicare Access and CHIP Reauthorization Act (MACRA) of 2015. The contract name is "Physician Cost Measures and Patient Relationship Codes (PCMP)." The contract number is 75FCMC18D0015, Task Order 75FCMC19F0004.

1.2 Measure Name

Cataract Removal with Intraocular Lens (IOL) Implantation Episode-based Cost Measure

1.3 Type of Measure

Cost/Resource Use

1.4 Data

The study period is January 1, 2021 through December 31, 2021. All episodes ending during the study period that meet inclusion and exclusion criteria are included in testing. The measure is calculated with Medicare Parts A and B, administrative claims data, Long-Term Minimum Data Set, Medicare Enrollment Database. For testing purpose, other data sources are used, including the American Community Survey, Common Medicare Environment.

Testing results are presented at a testing volume threshold of 10 episodes for clinician groups and individual practitioners. Clinician groups are identified by a Tax Identification Number (TIN). Individual clinicians are identified using a combination of a Tax Identification Number and National Provider Identifier (TIN-NPI).

[1]These documents will be available on the MACRA Feedback Page once field testing begins. https://www.cms.gov/Medicare/Quality-Payment-Program/Quality-Payment-Program/Give-Feedback

2.0 Preliminary Testing Results

This section presents preliminary testing results based on the revised measure as specified for the public comment period. Section 2.1 provides an overview of changes in the measure coverage, clinician population, and reliability between the current measure and the revised version of the measure. Sections 2.2 and 2.3 show additional evidence of scientific acceptability of the measure. Section 2.4 presents empirical results of the risk adjustment and stratification methods used by this measure. Section 2.5 examines the impact of adding social risk factors to the measure's risk adjustment model. Lastly, Section 2.6 examines the impact of exclusion criteria used by the measure through their frequency and resource use patterns.

2.1 Impacts of Revisions to the Measure

2.1.1 Measure Coverage

Table 1 shows the number of beneficiaries covered by this measure. Table 2 shows the characteristics of TINs and TIN-NPIs who are attributed at least 10 episodes. Compared to the current MIPS version, the revised version captures more beneficiaries.

Table 1: Measure Coverage

Metric	Value	
	Current MIPS Measure	Revised Measure
Number of Beneficiaries	416,509	715,350
Mean Age	73.68	74.15
Female %	61.79%	60.28%

Table 2: Clinician Characteristics

Metric	TIN				TIN-NPI			
	Current MIPS Measure		Revised Measure		Current MIPS Measure		Revised Measure	
	Count	%	Count	%	Count	%	Count	%
Count	3,795	100%	4,417	100%	7,642	100%	8,837	100%
Number of Episodes Attributed	-	-	-	-	-	-	-	-
20-39 Episodes	780	20.55%	669	15.15%	2,190	28.66%	1,826	20.66%
40-59 Episodes	562	14.81	475	10.75%	1,380	18.06%	1,351	15.29%
60-79 Episodes	347	9.14%	382	8.65%	823	10.77%	1,093	12.37%
80-99 Episodes	273	7.19%	322	7.29%	520	6.80%	778	8.80%
100-199 Episodes	622	16.39%	870	19.70%	920	12.04%	1,881	21.29%
200-299 Episodes	257	6.77%	372	8.42%	155	2.03%	443	5.01%

Metric	TIN				TIN-NPI			
	Current MIPS Measure		Revised Measure		Current MIPS Measure		Revised Measure	
	Count	%	Count	%	Count	%	Count	%
300+ Episodes	337	8.88%	697	15.78%	65	0.85%	276	3.12%
Census Region	-	-	-	-	-	-	-	-
Northeast	809	21.32%	917	20.76%	1,512	19.79%	1,808	20.46%
Midwest	779	20.53%	838	18.97%	1,711	22.39%	1,945	22.01%
South	1,326	34.94%	1,463	33.12%	2,748	35.96%	3,115	35.25%
West	858	22.61%	956	21.64%	1,647	21.55%	1,921	21.74%
Unknown	23	0.61%	43	0.97%	24	0.31%	48	0.54%

2.1.2 Frequently Attributed Specialties

Table 3 shows the most frequently attributed specialty for this measure, using a 10-episode testing volume threshold. The most frequently attributed specialty, ophthalmology, reflects the intent of the measure to capture costs of cataract removal with IOL implantation. Table 3 shows that Ophthalmology alone makes up over 99% of all clinicians who meet the testing volume threshold for the current MIPS and revised versions of the measure (99.9%).

Table 3: Count of the Top 6 Attributed Specialties

Current MIPS Version		Revised Version	
Specialty	Number of TIN-NPIs Attributed	Specialty	Number of TIN-NPIs Attributed
Ophthalmology	7,636	Ophthalmology	8,831

2.1.3 Reliability

Reliability evaluates a measure's ability to consistently differentiate the performance of one clinician from another. The signal-to-noise ratio is used to estimate reliability, which indicates how much of the variation in the measure score is explained by differences among clinician performance (i.e., signal) instead of differences within each clinician's performance (i.e., noise). Specifically, noise is the variation from one episode to another during the performance period for a particular clinician.

Table 4 shows reliability metrics at various testing volume thresholds. While higher thresholds yield higher reliability results, it is at the cost of further reducing the number of clinicians and clinician groups eligible for the measure, which would reduce the potential impact of the measure. For the purposes of testing, we used a 10-episode volume threshold (bolded in the table below) to align with the current MIPS version. If the measure is implemented in the MIPS in the future, CMS will establish a case minimum through notice-and-comment rulemaking.

Table 4: Sample Size, Mean Reliability, and Proportion of Clinicians above Moderate Reliability at Various Testing Volume Thresholds

Version	Testing Volume Threshold	TIN			TIN-NPI		
		Number of TINs	Mean Reliability	Percent Above 0.4	Number TIN-NPIs	Mean Reliability	Percent Above 0.4
Current Measure	**10**	**3,795**	**096**	**100.0%**	**7,642**	**0.95**	**100.0%**
Revised Measure	**10**	**4,217**	**0.96**	**100.0%**	**8,837**	**0.95**	**100.0%**
Current Measure	20	3,178	0.97	100.0%	6,053	0.96	100.0%
Revised Measure	20	3,787	0.97	100.0%	7,648	0.96	100.0%
Current Measure	30	2,727	0.98	100.0%	4,819	0.97	100.0%
Revised Measure	30	3,436	0.98	100.0%	6,679	0.97	100.0%

The results show that the revised measure can increase the number of clinicians participating in the measure by 1,195 without compromising the measure's reliability. At the testing volume of 10 episodes, the revised version shows very high reliability, specifically 0.96 at the TIN level and 0.95 at the TIN-NPI level (Table 4). CMS generally considers 0.4 as the threshold indicating 'moderate' reliability and 0.7 indicating 'high' reliability, which is supported by previous work into reliability and the threshold was finalized in the 2022 Physician Fee Schedule final rule.[2,3] All TINs and TIN-NPIs meet or exceed the moderate reliability threshold of 0.4 at the 10-episode testing volume threshold, 100.0% at both the TIN level and the TIN-NPI level.

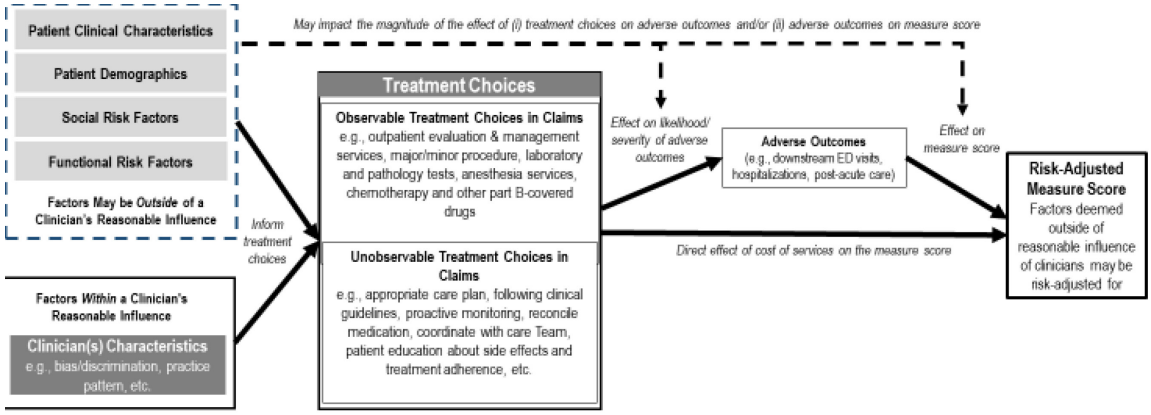
2.2 Validity

Validity is a criterion that evaluates whether the cost measure is able to quantify the construct that it aims to measure, which is the cost directly related to treatment choices and cost of adverse outcomes as a result of care. Validity is evaluated empirically by estimating the effect of relevant treatment choices on the measure score using multiple regression, based on the conceptual model outlined in Figure 1.

[2] 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

[3] CMS, "Medicare Program; CY 2022 Payment Policies Under the Physician Fee Schedule and Other Changes to Part B Payment Policies; Medicare Shared Savings Program Requirements; Provider Enrollment Regulation Updates; and Provider and Supplier Prepayment and Post-Payment Medical Review Requirements," 86 FR 64996-66031

Figure 1: Conceptual Model of the Relationship between Treatment Choices and the Measure Score



The cost measure is designed to reflect the cost directly related to treatment choices, as well as the cost of adverse outcomes as a result of care. Therefore, treatment choices, either observable in claims or otherwise, by an attributed clinician can directly impact the measure score or indirectly when they're mediated through the cost of adverse outcomes. The cost of adverse outcomes, in turn, contributes to the total costs that are captured by the measure score.

To demonstrate that the measure score is reflective of both the direct and indirect effects of treatment choices, this analysis first estimates the association between treatment choices and the measure score while controlling for the cost of adverse outcomes. Then, the association between treatment choices and the cost of adverse outcomes is estimated to demonstrate the indirect effect.

Generally, adverse outcomes are non-trigger inpatient hospitalizations, non-trigger emergency room visits, and post-acute care. The remaining service categories are generally considered treatment. For each of these categories, the regression models use the mean cost across episodes that were attributed to an individual clinician. The measure score is represented by a clinician's mean observed cost over expected cost ratio across their attributed episodes.

Overall, the results demonstrate that the cost measure is reflective of both the cost directly related to treatment choices, as well as the cost of adverse outcomes as a result of care (Table 5). Costs of adverse outcomes as a result of care were generally not statistically significant enough to be reflected by the measure.

Model 1 shows that the cost of adverse events is not associated with the measure score, which includes hospitalizations, emergency department visits, or post-acute care that are clinically related to cataract removal with IOL. The costs of outpatient evaluation and management, major or minor procedures, laboratory testing, anesthesia, and other services not classified elsewhere are associated with a worse measure score. This pattern suggests that, after controlling for the cost of adverse outcomes, these services remain to be cost drivers of cataract removal episodes. Model 2 shows that none of these services are associated with the cost of adverse events, therefore, it suggests that improvement in all these services can be made without negatively impacting the risk of adverse events.

Table 5: Estimated Effect of Treatment Choices (Revised Measure)

Categories of Service	Coefficient in Thousands [95% Confidence Interval] (p-value)			
	TIN		TIN-NPI	
	Model 1: Mean O/E = Mean Cost of Treatment Choices + Mean Cost of Adverse Events	Model 2: Mean Cost of Adverse Events = Mean Cost of Treatment Choices	Model 1: Mean O/E = Mean Cost of Treatment Choices + Mean Cost of Adverse Events	Model 2: Mean Cost of Adverse Events = Mean Cost of Treatment Choices
Adverse Events	-0.27 [-1.16,0.63] (p = 0.56)	-	-0.14 [-0.64,0.35] (p = 0.57)	-
Outpatient Evaluation & Management Services	0.22 [0.17,0.27] (p < 0.01)	0.00 [0.00,0.00] (p = 0.08)	0.22 [0.19,0.25] (p < 0.01)	0.00 [0.00,0.00] (p = 0.71)
Major Procedures	0.12 [0.03,0.21] (p < 0.01)	0.00 [0.00,0.00] (p = 0.48)	0.07 [0.00,0.14] (p = 0.04)	0.00 [0.00,0.00] (p = 0.78)
Ambulatory/Minor Procedures	0.06 [0.05,0.06] (p < 0.01)	0.00 [0.00,0.00] (p = 0.57)	0.06 [0.06,0.06] (p < 0.01)	0.00 [0.00,0.00] (p = 0.03)
Laboratory, Pathology, and Other Tests	0.56 [0.16,0.97] (p < 0.01)	-0.01 [-0.02,0.01] (p = 0.22)	0.67 [0.40,0.94] (p < 0.01)	0.00 [-0.01,0.01] (p = 0.56)
Anesthesia Services	0.14 [0.08,0.21] (p < 0.01)	0.00 [0.00,0.00] (p = 0.69)	0.22 [0.17,0.26] (p < 0.01)	0.00 [0.00,0.00] (p = 0.94)
Chemotherapy and Other Part B-Covered Drugs	0.26 [0.23,0.29] (p < 0.01)	0.00 [0.00,0.00] (p = 0.40)	0.29 [0.27,0.30] (p < 0.01)	0.00 [0.00,0.00] (p = 0.25)
All Other Services Not Otherwise Classified	0.14 [0.04,0.25] (p < 0.01)	0.00 [-0.01,0.00] (p = 0.41)	0.15 [0.06,0.23] (p < 0.01)	0.00 [0.00,0.00] (p = 0.83)

2.3 Performance Gap

Table 6 shows the distribution of the revised measure scores for clinicians and clinician groups. These results align with expectations based on our review of the literature and demonstrate that there is a performance gap in cost measure performance at both the clinician and clinician group levels. The Cataract Removal with Intraocular Lens (IOL) Implantation cost measure score at the 90th percentile is over 120% and 150% greater than the measure score at the 10th percentile at the TIN and TIN-NPI levels, respectively. The variation in the measure score, indicated by the interquartile range and standard deviation, is in the hundreds of dollars. The results suggest that there is opportunity for improvement in performance across providers.

Table 6: Distribution of the Measure Score (Revised Measure)

Metric	TIN	TIN-NPI
Mean Score	$3,122	$3,101
Score Interquartile Range (IQR)	$283	$288
Standard Deviation	$361	$355
Coefficient of Variation	0.12	0.11
Score Percentile		
10th	$2,833	$2,810
25th	$2,953	$2,942
50th	$3,068	$3,058
75th	$3,236	$3,230
90th	$3,473	$3,443

2.4 Risk Adjustment and Stratification

Figure 1 shows the conceptual model that outlines how patient-level and clinician-level factors can influence the measure score, which is informed by both published external research and our own data analysis.[4,5,6,7,8] The conceptual model includes risk factors that are either known by the literature or informed by the initial and reconvened Clinical Expert Workgroups to be within or outside of the influence of the attributed clinician. Risk factors, including social risk factors (SRFs), can both influence the treatment choices and impact the size of the effect of treatment choices by mitigating the risk of adverse outcomes and the cost of adverse outcomes.

A systematic approach then guides the decision of which factors to include in the risk adjustment model. First, during initial development of the current MIPS measure, we reviewed the literature to gather known risk factors and drivers of resource use. These factors are usually diagnoses; therefore, the first set of risk adjustors are commonly the Hierarchical Condition Categories. Then, we consulted our clinical expert panels on additional factors that are known to be associated with resource use. Together with our clinical expert panel, we reviewed the stratified results on episode cost across many different patient characteristics. We arrived at the final list of risk adjustors used in the current MIPS measure based on those discussions and consensus among the clinical experts. We also reviewed literature and gathered additional input from the reconvened Clinician Expert Workgroup to determine whether revisions should be made to the risk adjustment model. Additionally, during our testing phases, we also follow a structured and systematic approach to decide whether SRFs should be risk-adjusted for, which is further described in Section 2.5.

[4]Centers for Medicare & Medicaid (CMS), Office of Minority Health. "Utilization of Z Codes for Social Determinants of Health among Medicare Fee-for-Service Beneficiaries." (2019) https://www.cms.gov/files/document/z-codes-data-highlight.pdf
[5]Assistant Secretary of Health and Human Services for Planning and Evaluation. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. Washington, D.C. December 2016.
[6]Chen LM, Epstein AM, Orav EJ, Filice CE, Samson LW, Joynt Maddox KE. Association of Practice-Level Social and Medical Risk With Performance in the Medicare Physician Value-Based Payment Modifier Program. JAMA. 2017;318(5):453-461
[7]Medicare Payment Advisory Commission. Beneficiaries Dually Eligible for Medicare and Medicaid. 2018; https://www.macpac.gov/publication/data-book-beneficiaries-dually-eligible-for-medicare-and-medicaid-3/
[8]Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. https://aspe.hhs.gov/reports/second-report-congress-social-risk-medicares-value-based-purchasing-programs

2.4.1 Discrimination

Discrimination is a statistical criterion that evaluates the measure's ability to distinguish high-cost episodes from low-cost episodes, or the ability to explain the variance in cost of individual episodes. The amount of variance explained is estimated by the R-squared metric with the range between 0 and 1. The R-square value for the measure is 0.706 before and after adjusting for the model's complexity based on the number of risk adjustors used. In other words, 70.6% of the variation in the actual observed cost of episodes is explained by the risk adjustment model and sub-group stratification.

The remaining unexplained variance is due to variation in factors that are not adjusted for by the measure, such as the clinician's performance. The objective of a cost measure is to evaluate and differentiate the performance of clinicians. Therefore, achieving high explained variance is not essential because not all of the variation in cost of care should be adjusted. In collaboration with the experts from our clinical workgroup, this measure only adjusts for factors that are deemed to be outside of the influence of clinicians. Please see the Draft Cost Measure Methodology for more information on the full list of risk adjustors and sub-groups.

2.4.2 Calibration

Calibration evaluates the consistency of the measure in estimating episode cost across the full range of resource use patterns in the population. Calibration is estimated by the average predictive ratios across groups within the population, specifically groups are partitioned by deciles of expected episode cost. The predictive ratio is calculated using the formula of average expected cost / average observed cost for all episodes in each decile. A well-calibrated measure should have predictive ratios close to 1.00 across all deciles. In other words, such results show that the measure is consistent because it does not under- or over-predict cost throughout the range of resource use patterns in the population.

Table 7 shows that the model has consistent predictive ratios across risk score deciles, with each decile having a predictive ratio between 0.99 and 1.01. The average predictive ratio for all risk deciles is 1.00, which demonstrates that the risk adjustment does not under- or over- predict across the full range of resource use patterns in the population.

Table 7: Predictive Ratio by Decile of Predicted Episode Cost (Revised Measure)

Decile	Average Predictive Ratio
Decile 1	1.00
Decile 2	1.00
Decile 3	1.00
Decile 4	1.01
Decile 5	1.00
Decile 6	1.00
Decile 7	0.99
Decile 8	1.00
Decile 9	1.00
Decile 10	1.00

2.5 Social Risk Factor Analysis

Beyond clinical characteristics of patients, the cost of care may be influenced by non-clinical factors related to a patient's social risk factors (SRFs), such as race, income, education, and

employment. At the program level, MIPS adjusts for SRFs using the MIPS Complex Patient Bonus to ensure clinicians or groups treating more complex patients are not disadvantaged.[9] At the measure-level, the testing helps to navigate the tension between ensuring fairness for clinicians treating higher shares of vulnerable patients and the possibility of masking poor performance and perpetuating disparity if clinicians are held to different standards.

Table 8 outlines variables that may indicate SRFs and their advantages and disadvantages as indicators of individual-level SRFs. Based on availability of data, this analysis tested all variables except for the ICD-10 Z codes.

Table 8: Social Risk Factors Available for Analysis (Revised Measure)

Variable	Advantages	Disadvantages	Used in Testing
Dual Medicare and Medicaid enrollment status	• Available for all beneficiaries • Most powerful predictor of poor outcomes[10]	• Variation in Medicaid eligibility across states	Yes
Race/Ethnicity	• Available for most beneficiaries, except for ambiguous categories of "Unknown" or "Other"	• Social risk driven by someone's race is often correlated with and partially captured by dual status[11] • Only 5 categories available, which may lack granularity to fully capture disparities[12,13]	Yes
ICD-10 Z codes for social determinants of health	• Reflects individual-level factors that influence health status and contact with health services	• Not routinely and consistently coded on claims, only available for 0.1% of all fee-for-service claims in 2019[14]	No
American Community Survey	• Can link beneficiary's ZIP code to socioeconomic (SES) measurement of their neighborhood • Many SES indices can be derived from the survey data (e.g., Agency for Healthcare Research and Quality SES index, deprivation index)	• Only a proxy measure, not always accurate at individual-level	Yes

[9] https://qpp-cm-prod-content.s3.amazonaws.com/uploads/966/QPP%20COVID-19%20Response%20Fact%20Sheet.pdf
[10] Refer to footnote 4.
[11] Refer to footnote 4.
[12] Nguyen, Kevin H., Kaitlyn P. Lew, and Amal N. Trivedi. "Trends in Collection of Disaggregated Asian American, Native Hawaiian, and Pacific Islander Data: Opportunities in Federal Health Surveys." *American Journal of Public Health* (2022).
[13] Kader, Farah, Lan N. Doan, Matthew Lee, Matthew K. Chin, Simona C. Kwon, and Stella S. Yi. "Disaggregating Race/Ethnicity Data Categories: Criticisms, Dangers, And Opposing Viewpoints", *Health Affairs Forefront* (2022).
[14] Centers for Medicare & Medicaid (CMS), Office of Minority Health. "Utilization of Z Codes for Social Determinants of Health among Medicare Fee-for-Service Beneficiaries." (2019) https://www.cms.gov/files/document/z-codes-data-highlight.pdf

First, this analysis evaluated each of the variables for their association with episode cost using step-wise regression. In these models, race is a categorical variable (reference = White race). Table 9 shows that dual Medicare and Medicaid enrollment status is a consistent predictor, even in the presence of other variables. While race and AHRQ SES index are also shown to be associated with resource use, the literature suggests that dual status is still the best proxy of SRFs in predicting health outcomes based on its conceptual and empirical validity.[15]

Table 9: Associations of Available Social Risk Factor Variables and Cost of Care - TIN Reporting Level

Subgroup Risk Model	Variable	Coefficient (standard deviation, p-value)		
		Model 1: Base Model + Dual Status	Model 2: Base Model + Dual Status + Race	Model 3: Base Model + Dual Status + Race + AHRQ SES
ASC/ Bilateral	Dual Status	-$23.05 (SD: 4.51, p: <0.001)	-$11.28 (SD: 4.62, p: 0.01)	-$19.98 (SD: 4.65, p: <0.001)
	Race – Asian	-	-$46.02 (SD: 9.54, p: <0.001)	-$35.79 (SD: 9.61, p: <0.001)
	Race – Black	-	-$126.77 (SD: 5.35, p: <0.001)	-$133.73 (SD: 5.37, p: <0.001)
	Race – Hispanic	-	-$70.82 (SD: 11.73, p: <0.001)	-$76.97 (SD: 11.74, p: <0.001)
	Race – North American Native	-	$156.73 (SD: 14.71, p: <0.001)	$144.5 (SD: 14.72, p: <0.001)
	Race – Others	-	-$19.48 (SD: 5.37, p: <0.001)	-$12.82 (SD: 5.39, p: 0.02)
	Race –White	-	ref	ref
	AHRQ SES Index	-	-	-$3.3 (SD: 0.19, p: <0.001)
ASC / Unilateral	Dual Status	-$14.17 (SD: 3.5, p: <0.001)	-$13.3 (SD: 3.66, p: <0.001)	-$12.54 (SD: 3.69, p: <0.001)
	Race – Asian	-	$4.42 (SD: 6.7, p: 0.51)	$7.39 (SD: 6.76, p: 0.27)
	Race – Black	-	-$89.22 (SD: 4.14, p: <0.001)	-$84.57 (SD: 4.17, p: <0.001)
	Race – Hispanic	-	$5.67 (SD: 8.56, p: 0.51)	$8 (SD: 8.59, p: 0.35)
	Race – North American Native	-	$227.56 (SD: 13.1, p: <0.001)	$230.37 (SD: 13.12, p: <0.001)
	Race – Others	-	-$1.08 (SD: 4.68, p: 0.82)	-$1.54 (SD: 4.7, p: 0.74)
	Race – White	-	ref	ref
	AHRQ SES Index	-	-	$0.6 (SD: 0.17, p: <0.001)
HOPD / Bilateral	Dual Status	-$45.84 (SD: 14.12, p: <0.001)	-$16.37 (SD: 14.35, p: 0.25)	-$30.01 (SD: 14.43, p: 0.04)

[15] Office of the Assistant Secretary for Planning and Evaluation. "Second report to Congress on social risk and Medicare's value-based purchasing programs." (2020) https://aspe.hhs.gov/pdf-report/second-impact-report-to-congress

Subgroup Risk Model	Variable	Coefficient (standard deviation, p-value)		
		Model 1: Base Model + Dual Status	Model 2: Base Model + Dual Status + Race	Model 3: Base Model + Dual Status + Race + AHRQ SES
HOPD / Bilateral	Race – Asian	-	-$129.63 (SD: 38.39, p: <0.001)	-$113.45 (SD: 38.49, p: <0.001)
	Race – Black	-	-$361.29 (SD: 19.39, p: <0.001)	-$376.22 (SD: 19.46, p: <0.001)
	Race – Hispanic	-	-$130.95 (SD: 42.8, p: <0.001)	-$142.54 (SD: 42.86, p: <0.001)
	Race – North American Native	-	$42.97 (SD: 47.49, p: 0.37)	$26.02 (SD: 47.51, p: 0.58)
	Race – Others	-	-$69.98 (SD: 20.65, p: <0.001)	-$55.26 (SD: 20.71, p: 0.01)
	Race – White	-	ref	ref
	AHRQ SES Index	-	-	-$6.53 (SD: 0.73, p: <0.001)
HOPD / Unilateral	Dual Status	-$40.53 (SD: 8.03, p: <0.001)	-$12.74 (SD: 8.32, p: 0.13)	-$21.78 (SD: 8.38, p: 0.01)
	Race – Asian	-	-$95.76 (SD: 18.65, p: <0.001)	-$88.29 (SD: 18.72, p: <0.001)
	Race – Black	-	-$228.81 (SD: 10.02, p: <0.001)	-$233.63 (SD: 10.09, p: <0.001)
	Race – Hispanic	-	-$55.88 (SD: 21.25, p: 0.01)	-$57.81 (SD: 21.3, p: 0.01)
	Race – North American Native	-	$78.37 (SD: 33.11, p: 0.02)	$66.12 (SD: 33.12, p: 0.05)
	Race – Others	-	-$41.26 (SD: 13.13, p: <0.001)	-$37.13 (SD: 13.16, p: <0.001)
	Race – White	-	ref	ref
	AHRQ SES Index	-	-	-$3.73 (SD: 0.45, p: <0.001)

The subsequent analyses focus on dual status as the main proxy variable for SRFs for risk adjustment. To determine whether it's appropriate to risk adjust for SRFs, the following criteria are considered:

(i) whether there's an association between social risk and performance by examining the coefficient of patient-level dual status when added into the risk model,

(ii) whether the observed association is most influenced by patient-level factors or clinician-level factors by examining the stability of the patient-level dual status coefficient after adding clinician's dual share variable, as well as including the clinician's fixed effects,

(iii) whether the patient's need or complexity (rather than poor quality) is driving the observed performance differences by examining the differences in performance on dual patients versus non-dual patients and if there are many clinicians who are able to perform similarly or better on their dual patients than their non-dual patients, and

(iv) the impact of risk adjusting for SRFs by examining the performance shift of clinicians compared to a risk adjustment model that doesn't risk adjust for SRFs.

There's a statistically significant negative association between the patient's dual status and episode cost (Table 10). This association fluctuates slightly but remains statistically significant after adding variables to account for provider-level factors, which suggests that the patient-level factors and provider-level factors are both influential. There is no performance degradation observed with increasing share of dual episodes (Table 11). There are more clinicians who are able to perform significantly better on their dual episodes than their non-dual episodes and there are many who are able to perform equally well on their dual episodes as their non-dual episodes, which suggests that it is possible to mitigate the effect of SRFs (Table 12). Lastly, risk adjusting for dual status does not appear to substantially change the performance ranking for many providers (Table 13).

Table 10: Coefficient of Patient-level Dual Status under Different Models (Revised Measure)

Level	Subgroup Risk Model	% of All Episodes	Coefficient of Patient-level Dual Status (standard deviation, p-value)		
			Base Model + Patient-level Dual Status	Base Model + Patient-level Dual Status + Clinician's Dual Share	Base Model + Patient-level Dual Status + Clinician's Fixed Effect
TIN	ASC / Bilateral	369,891 (40.45%)	-23.05 (SD: 4.51, p: <0.001)	-47.54 (SD: 4.68, p: <0.001)	-44.13 (SD: 3.94, p: <0.001)
TIN	ASC / Unilateral	338,184 (36.98%)	-14.17 (SD: 3.5, p: <0.001)	-36.11 (SD: 3.79, p: <0.001)	-33.67 (SD: 3.44, p: <0.001)
TIN	HOPD / Bilateral	89,358 (9.77%)	-45.84 (SD: 14.12, p: <0.001)	-33.01 (SD: 14.59, p: 0.02)	-38.11 (SD: 13.04, p: <0.001)
TIN	HOPD / Unilateral	116,965 (12.79%)	-40.53 (SD: 8.03, p: <0.001)	-18.15 (SD: 8.56, p: 0.03)	-9.59 (SD: 8.03, p: 0.23)
TIN-NPI	ASC / Bilateral	369,919 (40.45%)	-23.04 (SD: 4.51, p: <0.001)	-46.01 (SD: 4.71, p: <0.001)	-45.04 (SD: 3.89, p: <0.001)
TIN-NPI	ASC / Unilateral	338,192 (36.98%)	-14.2 (SD: 3.5, p: <0.001)	-34.84 (SD: 3.83, p: <0.001)	-31.07 (SD: 3.44, p: <0.001)
TIN-NPI	HOPD / Bilateral	89,370 (9.77%)	-45.73 (SD: 14.12, p: <0.001)	-28.82 (SD: 14.74, p: 0.05)	-27.15 (SD: 12.92, p: 0.04)
TIN-NPI	HOPD / Unilateral	116,968 (12.79%)	-40.59 (SD: 8.03, p: <0.001)	-19.68 (SD: 8.72, p: 0.02)	-7.81 (SD: 7.96, p: 0.33)

Table 11: Mean Ratio of Observed Cost to Expected Cost (O/E) Stratified by Clinician's Dual Share and Patient's Dual Status (Revised Measure)

Dual Share	TIN			TIN-NPI		
	All Episodes	Dual Episodes	Non-Dual Episodes	All Episodes	Dual Episodes	Non-Dual Episodes
All	1.01	1.00	1.01	1.00	0.99	1.00
0%	1.00	-	1.00	1.00	-	1.00
1-20%	1.01	1.00	1.01	1.00	0.99	1.00
21-40%	1.01	1.00	1.01	1.01	1.01	1.01
41-60%	1.00	1.00	1.01	1.01	1.01	1.01
61-80%	1.00	1.00	1.01	0.99	0.99	0.99
81-99%	0.97	0.97	0.98	0.97	0.97	0.97
100%	1.00	1.00	-	0.99	0.99	-

Table 12: Proportions of Clinicians Who Perform Significantly Worse, Equally Well, or Significantly Better on Their Dual Episodes than Non-Dual Episodes (Revised Measure)

Reporting Level	Significantly Worse	Equally Well	Significantly Better
TIN	4%	88%	8%
TIN-NPI	3%	91%	7%

Table 13: Clinicians' Performance Shift Measured by the Change in the Average Ratio of Observed Cost to Expected Cost (O/E) (Revised Measure)

Reporting Level	Proportions of Clinicians Affected at Various Levels of Performance Shift	
	Ranking Shift by 1% or more	Ranking Shift by 5% or more
TIN	21.5%	0.3%
TIN-NPI	19.9%	0.3%

2.6 Impact of Exclusions

Table 14 displays descriptive statistics of all episodes meeting the revised measure's triggering logic, excluded episodes, and final reportable episodes at both TIN and TIN-NPI levels. These exclusion criteria ensure that the reportable episode populations are more homogenous and comparable than all episodes meeting triggering logic. It is worth noting that only the observed cost is shown, which has not been risk adjusted for using our risk adjustment model. Therefore, the differences in cost may appear much smaller after risk adjustment than as-is.

Overall, exclusion criteria decrease the distribution of observed cost of all episodes meeting trigger logic, from the mean of $3,110 to $3,103 at the TIN-level and $3,105 at the TIN-NPI level.

Episodes where the procedure was not performed in OP, IP, or ASC settings were excluded because these are the standard settings for this procedure. These episodes also have a higher mean observed cost than all episodes meeting triggering logic, at $3,174, likely because the procedure was done in an emergency or due to cataract complications.

Episodes where a beneficiary died before the episode end date are excluded because they do not provide sufficient data in the episode window period. These episodes also have a higher mean observed cost than all episodes meeting triggering logic, at $2,884, likely because the costs are distributed over fewer days than a typical episode.

Episodes that could not be classified into measure subgroups are excluded because they do not include sufficient procedure setting or intensity (unilateral vs bilateral) data. These episodes also have a higher mean observed cost than all episodes meeting triggering logic, at $2,917, likely because the costs are a composite amount of the four subgroups unclassified.

Episodes classified as outlier cases are excluded because they deviate substantially from the projected cost for a given patient risk profile. Outlier episodes have a mean observed episode cost of $3,673 compared to $3,110 for all episodes meeting triggering logic. The wide variability of observed episode costs for outlier cases also supports their exclusion. At the 10th percentile the outlier cases observed cost is $633 and at the 90th percentile the observed cost is $7,405.

Episodes where the TIN/TIN-NPI did not meet the case minimum are excluded because this indicates the procedure is an outlier for the clinician/clinician group. This is likely why the mean costs for these episodes is lower than all episodes meeting triggering logic.

Based on the input from the clinical expert workgroup, episodes with patients with significant ocular conditions impacting surgical complication rate/visual outcomes are excluded because these episodes can be clinically distinct from the overall cataract removal population. These episodes only have a slightly higher mean observed cost than all episodes meeting triggering logic, at $3,108, which suggests that cataract removal with IOL resource use patterns may need to be reevaluated to determine better care intensity stratifications.

Table 14: Cost Statistics for Measure Exclusions (Revised Measure)

Exclusion Criteria	Episodes		Observed Episode Cost					
				Percentile				
	Count	Percent of All Episodes Meeting Trigger Logic	Mean	10th	25th	50th	75th	90th
All Episodes Meeting Triggering Logic	903,542	100.00%	$3,110	$1,750	$1,984	$3,067	$3,724	$4,872
Not in OP, IP, or ASC Setting	53	0.01%	$3,174	$1,717	$1,969	$3,028	$3,713	$5,480
Beneficiary Death in Episode	4,177	0.46%	$2,884	$1,705	$1,900	$2,762	$3,601	$4,480
Episodes that Cannot be Classified into a Subgroup	5,054	0.56%	$2,917	$899	$1,740	$2,859	$3,695	$5,633
Outlier	15,824	1.75%	$3,673	$633	$873	$2,886	$5,949	$7,405
Patients with Significant Ocular Conditions Impacting Surgical Complication Rate/Visual Outcomes	104,092	11.52%	$3,108	$1,686	$1,961	$2,940	$3,736	$4,839
TIN does not Meet Case Minimum	6,391	0.71%	$2,835	$1,017	$1,838	$2,690	$3,581	$5,385
TIN-NPI does not Meet Case Minimum	15,292	1.69%	$2,887	$1,092	$1,885	$2,805	$3,596	$5,141
Reportable Episodes - Group Reporting	772,799	85.53%	$3,103	$1,776	$1,996	$3,098	$3,716	$4,655
Reportable Episodes - Individual Reporting	767,273	84.92%	$3,105	$1,777	$1,997	$3,105	$3,717	$4,657

Appendix A. Distributions of Measure Score (Revised Measure)

Figure 2: Distribution of Measure Score - TIN

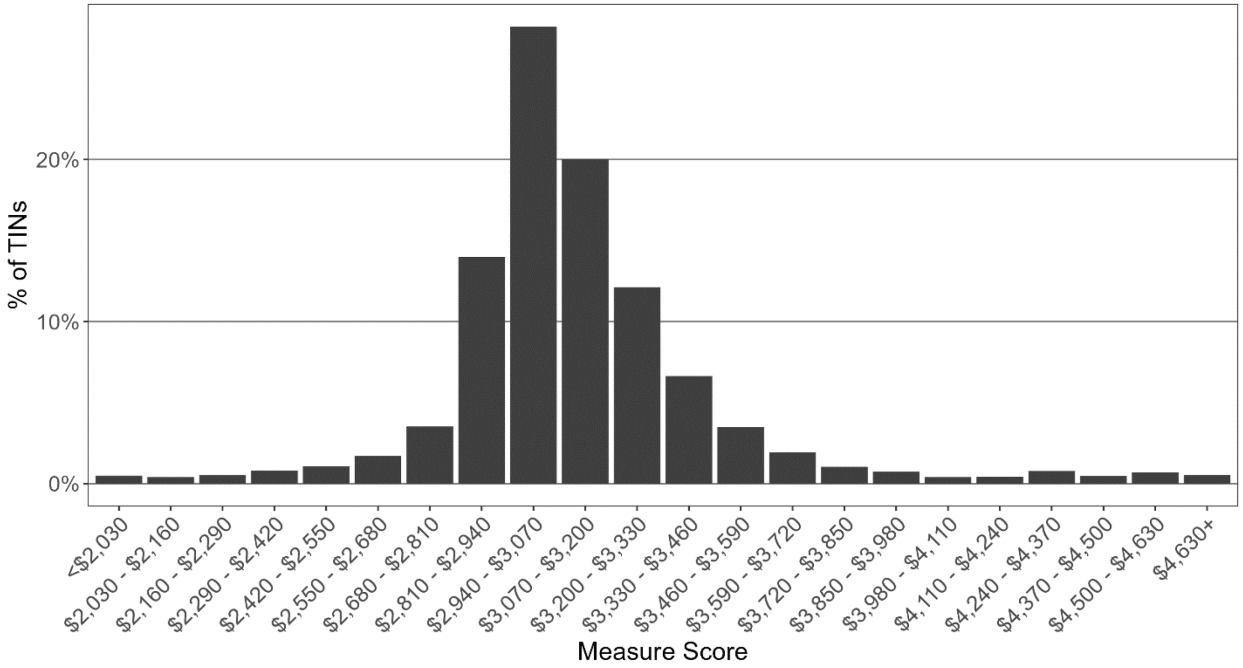
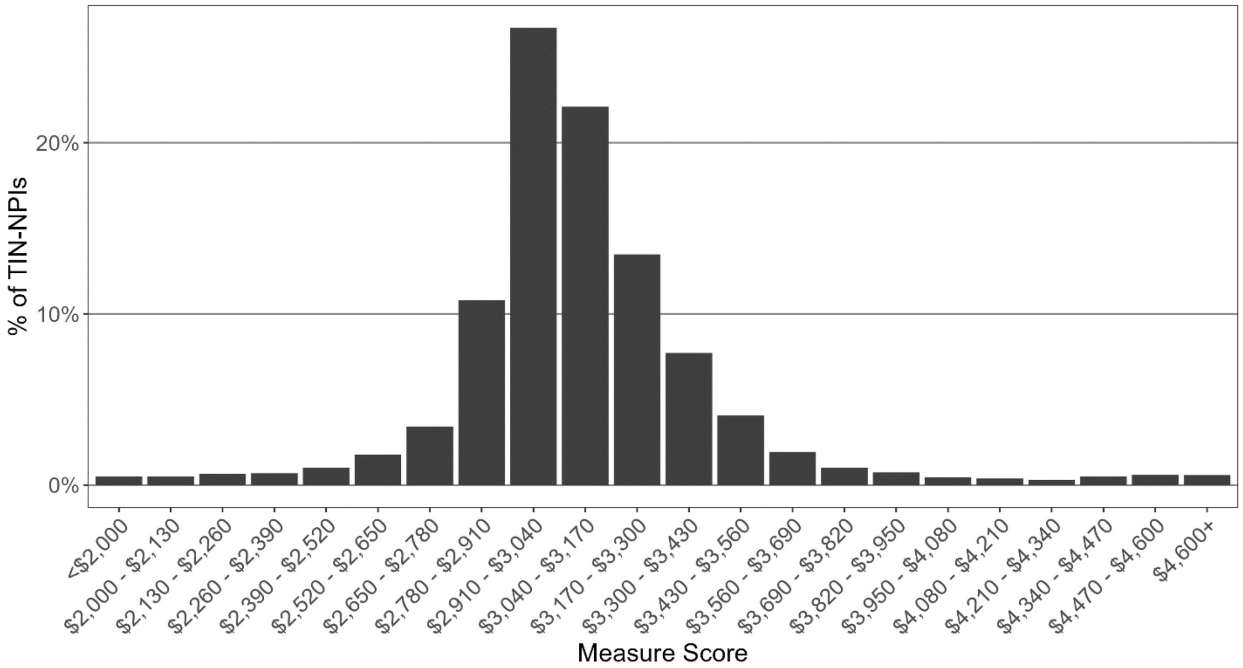


Figure 3: Distribution of Measure Score - TIN-NPI



Appendix B. Associations between Social Risk Factor Variables and Cost of Care for TIN-NPIs

Table B1: Associations of Available Social Risk Factor Variables and Cost of Care – TIN-NPI Reporting Level

Subgroup Risk Model	Variable	Coefficient (standard deviation, p-value)		
		Model 1: Base Model + Dual Status	Model 2: Base Model + Dual Status + Race	Model 3: Base Model + Dual Status + Race + AHRQ SES
ASC/ Bilateral	Dual Status	-$23.04 (SD: 4.51, p: <0.001)	-$11.27 (SD: 4.62, p: 0.01)	-$19.97 (SD: 4.65, p: <0.001)
	Race – Asian	-	-$46.01 (SD: 9.54, p: <0.001)	-$35.79 (SD: 9.61, p: <0.001)
	Race – Black	-	-$126.76 (SD: 5.35, p: <0.001)	-$133.73 (SD: 5.37, p: <0.001)
	Race – Hispanic	-	-$70.82 (SD: 11.73, p: <0.001)	-$76.96 (SD: 11.74, p: <0.001)
	Race – North American Native	-	$156.73 (SD: 14.71, p: <0.001)	$144.5 (SD: 14.72, p: <0.001)
	Race – Others	-	-$19.48 (SD: 5.37, p: <0.001)	-$12.82 (SD: 5.39, p: 0.02)
	Race –White	-	ref	ref
	AHRQ SES Index	-	-	-$3.3 (SD: 0.19, p: <0.001)
ASC / Unilateral	Dual Status	-$14.2 (SD: 3.5, p: <0.001)	-$13.34 (SD: 3.66, p: <0.001)	-$12.57 (SD: 3.69, p: <0.001)
	Race – Asian	-	$4.43 (SD: 6.7, p: 0.51)	$7.4 (SD: 6.76, p: 0.27)
	Race – Black	-	-$89.22 (SD: 4.14, p: <0.001)	-$84.56 (SD: 4.17, p: <0.001)
	Race – Hispanic	-	$5.69 (SD: 8.56, p: 0.51)	$8.03 (SD: 8.59, p: 0.35)
	Race – North American Native	-	$227.57 (SD: 13.1, p: <0.001)	$230.38 (SD: 13.12, p: <0.001)
	Race – Others	-	-$1.07 (SD: 4.68, p: 0.82)	-$1.54 (SD: 4.7, p: 0.74)
	Race – White	-	ref	ref
	AHRQ SES Index	-	-	$0.6 (SD: 0.17, p: <0.001)
HOPD / Bilateral	Dual Status	-$45.73 (SD: 14.12, p: <0.001)	-$16.26 (SD: 14.35, p: 0.26)	-$29.9 (SD: 14.43, p: 0.04)
	Race – Asian	-	-$129.66 (SD: 38.39, p: <0.001)	-$113.47 (SD: 38.49, p: <0.001)
	Race – Black	-	-$361.3 (SD: 19.39, p: <0.001)	-$376.24 (SD: 19.46, p: <0.001)
	Race – Hispanic	-	-$131 (SD: 42.8, p: <0.001)	-$142.59 (SD: 42.86, p: <0.001)
	Race – North American Native	-	$42.9 (SD: 47.49, p: 0.37)	$25.95 (SD: 47.51, p: 0.58)
	Race – Others	-	-$70.01 (SD: 20.65, p: <0.001)	-$55.28 (SD: 20.71, p: 0.01)

Subgroup Risk Model	Variable	Coefficient (standard deviation, p-value)		
		Model 1: Base Model + Dual Status	Model 2: Base Model + Dual Status + Race	Model 3: Base Model + Dual Status + Race + AHRQ SES
HOPD/ Bilateral	Race – White	-	ref	ref
	AHRQ SES Index	-	-	$-6.53 (SD: 0.73, p: <0.001)
HOPD / Unilateral	Dual Status	-$40.59 (SD: 8.03, p: <0.001)	-$12.82 (SD: 8.32, p: 0.12)	-$21.87 (SD: 8.38, p: 0.01)
	Race – Asian	-	-$95.71 (SD: 18.65, p: <0.001)	$-88.23 (SD: 18.72, p: <0.001)
	Race – Black	-	-$228.81 (SD: 10.02, p: <0.001)	-$233.64 (SD: 10.09, p: <0.001)
	Race – Hispanic	-	-$55.83 (SD: 21.25, p: 0.01)	-$57.77 (SD: 21.3, p: 0.01)
	Race – North American Native	-	$78.37 (SD: 33.11, p: 0.02)	$66.11 (SD: 33.12, p: 0.05)
	Race – Others	-	-$40.67 (SD: 13.13, p: <0.001)	-$36.54 (SD: 13.16, p: 0.01)
	Race – White	-	ref	ref
	AHRQ SES Index	-	-	-$3.73 (SD: 0.45, p: <0.001)