

Quality Payment PROGRAM

Rheumatoid Arthritis

Measure Testing Form

January 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	Measure Coverage.....	4
2.2	Frequently Attributed Specialties	5
2.3	Reliability	5
2.4	Validity	6
2.5	Performance Gap.....	8
2.6	Risk Adjustment and Stratification	8
	2.6.1 Discrimination.....	9
	2.6.2 Calibration	9
2.7	Social Risk Factor Analysis	10
2.8	Impact of Exclusions	15
	Appendix A. Distributions of Measure Score.....	17

1.0 Introduction

This Measure Testing Form (MTF) provides a brief summary of the preliminary measure testing results as part of field testing five episode-based cost measures. Readers may review these results, alongside other documentation, to provide feedback on the draft measure using the [field testing survey](#). The testing results reflect the performance of the measure as specified at the time of field testing, which is part of the measure development process. 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 care episode and patient condition groups for use in cost measures to meet the requirements of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). The contract name is “Physician Cost Measures and Patient Relationship Codes (PCMP).” The contract number is 75FCMC18D0015, Task Order 75FCMC19F0004.

1.2 Measure Name

Rheumatoid Arthritis 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, B, and D administrative claims data, Long-Term Minimum Data Set, Medicare Enrollment Database. For testing purpose, other data sources are used, including the American Community Survey and Common Medicare Environment.

Testing results are presented at a testing volume threshold of 20 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 measure as specified for field testing. Section 2.1 provides an overview of the measure's coverage of beneficiaries and cost. Section 2.2 lists the most frequently attributed specialties. Sections 2.3 through 2.5 provide evidence of scientific acceptability of the measure. Section 2.6 presents empirical results of the risk adjustment and stratification methods used by this measure. Section 2.7 examines the impact of adding social risk factors to the measure's risk adjustment model. Lastly, Section 2.8 examines the impact of exclusion criteria used by the measure through their frequency and resource use patterns.

2.1 Measure Coverage

Table 1 shows the number of patients and the amount of Medicare Parts A and B cost covered by this measure. Table 2 shows the characteristics of TINs and TIN-NPIs who are attributed at least 20 episodes. This measure has the potential to have high impact as rheumatoid arthritis is a common condition for the Medicare population, as shown in the results below.

Table 1: Measure Coverage

Metric	Value
Number of Beneficiaries	336,327
Mean Age	72.5
Female %	76.1%
Percent of Parts A & B Costs Covered by the Measure – TIN	0.79%
Percent of Parts A & B Costs Covered by the Measure – TIN-NPI	0.63%

Table 2: Clinician Characteristics

Metric	TIN		TIN-NPI	
	Count	%	Count	%
Count	3,342	100.0%	3,874	100.0%
Number of Episodes Attributed				
20-39 Episodes	1,167	34.9%	1,355	35.0%
40-59 Episodes	504	15.1%	837	21.6%
60-79 Episodes	310	9.3%	537	13.9%
80-99 Episodes	193	5.8%	402	10.4%
100-199 Episodes	621	18.6%	652	16.8%
200-299 Episodes	228	6.8%	84	2.2%
300+ Episodes	319	9.5%	7	0.2%
Census Region				
Northeast	604	18.1%	755	19.5%
Midwest	636	19.0%	830	21.4%
South	1,432	42.8%	1,540	39.8%
West	664	19.9%	743	19.2%
Unknown	6	0.2%	6	0.2%

2.2 Frequently Attributed Specialties

Table 3 shows the top 10 attributed specialties for this measure, using a 20-episode testing volume threshold. The most frequently attributed specialties reflect the intent of the measure to capture costs of the management of rheumatoid arthritis, including rheumatologists, primary care, and related specialists. These clinicians are also consistent with input provided by stakeholders, including patient and family partners (PFPs), during the measure development process. PFPs identified rheumatologists, physical and occupational therapists, primary care physicians, and behavioral health practitioners amongst others as being part of their care team.

Table 3: Count of the Top 10 Attributed Specialties

Specialty	Number of TIN-NPIs Attributed
Rheumatology	3,172
Internal Medicine	286
Nurse Practitioner	200
Physician Assistant	129
Family Practice	41
General Practice	7
Interventional Pain Management	6
Pain Management	6
Allergy/Immunology	5
Physical Medicine and Rehabilitation	3

2.3 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 clinicians 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 field testing, we used a 20-episode volume threshold (bolded in the table below). If the measure is implemented in the Merit-based Incentive Payment System (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

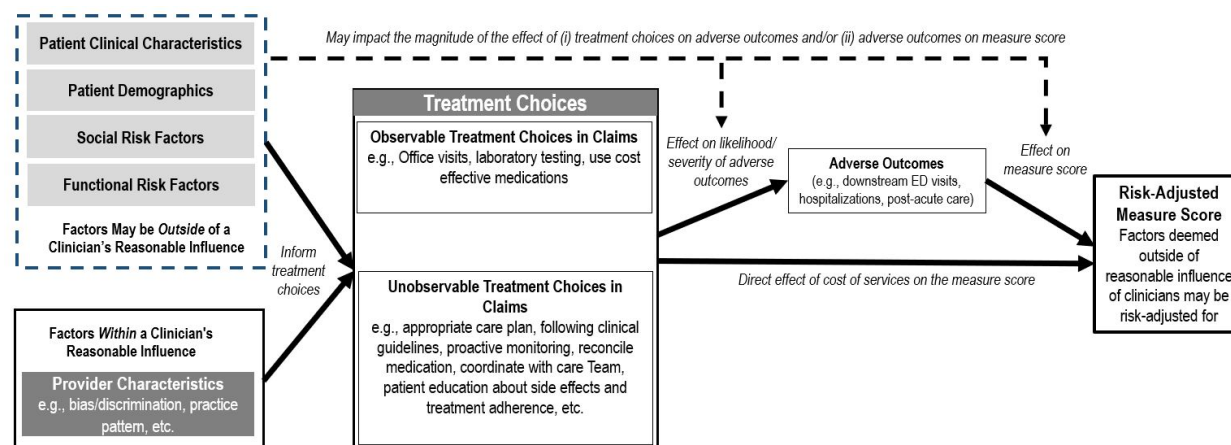
Testing Volume Threshold	TIN			TIN-NPI		
	Number of TINs	Mean Reliability	Percent Above 0.4	Number TIN-NPIs	Mean Reliability	Percent Above 0.4
10	5,753	0.596	75.51%	5,843	0.665	86.09%
20	3,342	0.712	93.72%	3,874	0.743	97.06%
30	2,588	0.762	98.45%	3,059	0.780	99.51%

At the testing volume of 20 episodes, the mean reliability for the Rheumatoid Arthritis measure is high, specifically 0.71 at the TIN level and 0.74 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} The vast majority of all TINs and TIN-NPIs meet or exceed the moderate reliability threshold of 0.4 at the 20-episode testing volume threshold, with 93.72% and 97.06% at both reporting levels, respectively.

2.4 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.

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 impact the measure score directly 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

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](https://www.federalregister.gov/documents/2021/08/16/2021-16496/medicare-program-cy-2022-payment-policies).

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 cost of adverse outcomes as a result of care (Table 5). Therefore, there's evidence that the measure is capturing what it purports to measure.

Model 1 shows that the cost of adverse events is associated with a worse measure score. Major procedures and medications from parts B and D are also associated with a worse measure score. However, model 2 shows that these services are associated with lower cost of adverse events at the TIN reporting level and the association is less clear at the TIN-NPI reporting level. On the other hand, outpatient evaluation and management services and durable medical equipment are associated with higher cost of adverse events, which may reflect higher clinical needs associated with complications.

Table 5: Estimated Effect of Treatment Choices

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	0.08 [0.07,0.08] (p < 0.01)		0.06 [0.05,0.07] (p < 0.01)	
Outpatient Evaluation & Management Services	0.01 [-0.02,0.04] (p = 0.45)	1.46 [1.30,1.61] (p < 0.01)	0.00 [-0.04,0.05] (p = 0.82)	0.45 [0.33,0.57] (p < 0.01)
Major Procedures	0.08 [0.04,0.12] (p < 0.01)	-0.13 [-0.32,0.06] (p = 0.19)	0.10 [0.06,0.13] (p < 0.01)	0.23 [0.12,0.34] (p < 0.01)
Durable Medical Equipment and Supplies	-0.04 [-0.16,0.07] (p = 0.46)	2.79 [2.21,3.37] (p < 0.01)	0.01 [-0.12,0.14] (p = 0.91)	1.09 [0.69,1.48] (p < 0.01)
Chemotherapy and Other Part B- Covered Drugs	0.07 [0.07,0.07] (p < 0.01)	-0.04 [-0.05,-0.03] (p < 0.01)	0.07 [0.07,0.08] (p < 0.01)	0.00 [-0.01,0.01] (p = 0.96)
Part-D Drugs	0.07 [0.06,0.07] (p < 0.01)	-0.03 [-0.04,-0.01] (p < 0.01)	0.07 [0.07,0.07] (p < 0.01)	0.02 [0.01,0.02] (p < 0.01)

2.5 Performance Gap

Table 6 shows the distribution of the measure score 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 Rheumatoid Arthritis cost measure score at the 90th percentile is over 2 times greater than the measure score at the 10th percentile at both the TIN and TIN-NPI levels. The variation in the measure score, indicated by the interquartile range and standard deviation, is in the thousands of dollars. The results suggest that there is opportunity for improvement in performance across clinicians.

Table 6: Distribution of the Measure Score

Metric	TIN	TIN-NPI
Mean Score	\$11,519	\$12,313
Score Interquartile Range (IQR)	\$4,929	\$6,329
Standard Deviation	\$4,203	\$5,028
Coefficient of Variation	0.36	0.41
Score Percentile		
10 th	\$6,690	\$6,448
25 th	\$8,732	\$8,812
50 th	\$11,112	\$11,793
75 th	\$13,661	\$15,142
90 th	\$16,805	\$18,689

2.6 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 Clinical Expert Workgroup 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, 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 adjusters are

⁴Office 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>.

⁵Assistant 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.

⁶Chen 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

⁷Medicare 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/>.

⁸Office 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-medicare-value-based-purchasing-programs>.

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 based on those discussions and consensus among the clinical experts. 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 2.7.

2.6.1 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 0.154, and 0.154 after adjusting for the model's complexity based on the number of risk adjustors used. In other words, 15.4% 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.

2.6.2 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 there is some under-prediction for the low-risk episodes and over-prediction for the high-risk episodes. The results suggest opportunities to improve the risk classification by refining the risk adjustors and further examining drivers of costs.

Table 7: Predictive Ratio by Decile of Predicted Episode Cost

Decile	Average Predictive Ratio
Decile 1	0.72
Decile 2	0.78
Decile 3	0.83
Decile 4	0.81
Decile 5	0.87

Decile	Average Predictive Ratio
Decile 6	0.93
Decile 7	0.97
Decile 8	1.01
Decile 9	1.07
Decile 10	1.33

2.7 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

Variable	Advantages	Disadvantages	Used in Testing
Dual Medicare and Medicaid enrollment status	<ul style="list-style-type: none"> Available for all beneficiaries Most powerful predictor of poor outcomes¹⁰ 	<ul style="list-style-type: none"> Variation in Medicaid eligibility across states 	Yes
Race/Ethnicity	<ul style="list-style-type: none"> Available for most beneficiaries, except for ambiguous categories of "Unknown" or "Other" 	<ul style="list-style-type: none"> Social risk driven by someone's race is often correlated with and partially captured by dual status¹¹ Only 5 categories available, which may lack granularity to fully capture disparities^{12, 13} 	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).

Variable	Advantages	Disadvantages	Used in Testing
ICD-10 Z codes for social determinants of health	<ul style="list-style-type: none"> Reflects individual-level factors that influence health status and contact with health services 	<ul style="list-style-type: none"> Not routinely and consistently coded on claims, only available for 0.1% of all fee-for-service claims in 2019¹⁴ 	No
American Community Survey	<ul style="list-style-type: none"> 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) 	<ul style="list-style-type: none"> Only a proxy measure, not always accurate at individual-level 	Yes

First, this analysis evaluated each of the variables for their association with episode cost using step-wise regression. As these are categorical variables, we used the Race-white variable as the reference variable. Table 9 shows that dual Medicare and Medicaid enrollment status is the most consistent predictor at the TIN level, even in the presence of other variables. This is also consistent with other research that found dual status to be the best proxy of SRFs in predicting health outcomes.¹⁵ However, at the TIN-NPI level, Asian race is the most important predictor, followed closely by dual status.

Table 9: Associations of Available Social Risk Factor Variables and Cost of Care

Level	Subgroup Risk Model	Variable	Coefficient in Log-transformed Episode Cost (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
TIN	With Part D Coverage	Dual Status	0.29 (SD: 0.01, p: <0.001)	0.31 (SD: 0.01, p: <0.001)	0.33 (SD: 0.01, p: <0.001)
		Race – Asian	-	-0.27 (SD: 0.02, p: <0.001)	-0.29 (SD: 0.02, p: <0.001)
		Race – Black	-	-0.11 (SD: 0.01, p: <0.001)	-0.09 (SD: 0.01, p: <0.001)
		Race – Hispanic	-	0.05 (SD: 0.02, p: <0.001)	0.07 (SD: 0.02, p: <0.001)
		Race – North American Native	-	0.16 (SD: 0.03, p: <0.001)	0.19 (SD: 0.03, p: <0.001)

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>

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>

Level	Subgroup Risk Model	Variable	Coefficient in Log-transformed Episode Cost (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
	Without Part D Coverage	Race – Others	-	-0.06 (SD: 0.02, p: <0.001)	-0.07 (SD: 0.02, p: <0.001)
		Race – White	-	ref	ref
		AHRQ SES Index	-	-	0.01 (SD: 0, p: <0.001)
		Dual Status	0.04 (SD: 0.07, p: 0.54)	0.04 (SD: 0.07, p: 0.54)	0.05 (SD: 0.07, p: 0.48)
		Race – Asian	-	-0.1 (SD: 0.04, p: 0.02)	-0.09 (SD: 0.04, p: 0.04)
		Race – Black	-	-0.02 (SD: 0.01, p: 0.11)	-0.02 (SD: 0.02, p: 0.28)
		Race – Hispanic	-	-0.16 (SD: 0.04, p: <0.001)	-0.15 (SD: 0.04, p: <0.001)
		Race – North American Native	-	0.21 (SD: 0.04, p: <0.001)	0.23 (SD: 0.04, p: <0.001)
		Race – Others	-	-0.08 (SD: 0.03, p: <0.001)	-0.08 (SD: 0.03, p: <0.001)
		Race – White	-	ref	ref
		AHRQ SES Index	-	-	0 (SD: 0, p: <0.001)
TIN-NPI	With Part D Coverage	Dual Status	0.33 (SD: 0.01, p: <0.001)	0.37 (SD: 0.01, p: <0.001)	0.39 (SD: 0.01, p: <0.001)
		Race – Asian	-	-0.38 (SD: 0.02, p: <0.001)	-0.4 (SD: 0.02, p: <0.001)
		Race – Black	-	-0.15 (SD: 0.01, p: <0.001)	-0.12 (SD: 0.01, p: <0.001)
		Race – Hispanic	-	0.06 (SD: 0.02, p: 0.01)	0.08 (SD: 0.02, p: <0.001)
		Race – North American Native	-	0.21 (SD: 0.04, p: <0.001)	0.24 (SD: 0.04, p: <0.001)
		Race – Others	-	-0.08 (SD: 0.02, p: <0.001)	-0.1 (SD: 0.02, p: <0.001)
		Race – White	-	ref	ref
		AHRQ SES Index	-	-	0.01 (SD: 0, p: <0.001)
	Without Part D Coverage	Dual Status	0.03 (SD: 0.09, p: 0.69)	0.03 (SD: 0.09, p: 0.7)	0.04 (SD: 0.09, p: 0.62)
		Race – Asian	-	-0.12 (SD: 0.05, p: 0.03)	-0.11 (SD: 0.05, p: 0.03)
		Race – Black	-	-0.01 (SD: 0.02, p: 0.59)	0 (SD: 0.02, p: 0.86)
		Race – Hispanic	-	-0.14 (SD: 0.05, p: 0.01)	-0.13 (SD: 0.05, p: 0.02)
		Race – North American Native	-	0.22 (SD: 0.05, p: <0.001)	0.24 (SD: 0.05, p: <0.001)
		Race – Others	-	-0.09 (SD: 0.03, p: 0.01)	-0.09 (SD: 0.03, p: 0.01)
		Race – White	-	ref	ref

Level	Subgroup Risk Model	Variable	Coefficient in Log-transformed Episode Cost (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
TIN-NPI		AHRQ SES Index	-	-	0 (SD: 0, 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 association between the patient's dual status and episode cost in the largest subgroup (Table 10). This association is relatively stable and remains statistically significant after adding variables to account for clinician-level factors, which suggests that the patient-level factors are more influential than clinician-level factors. This is also supported by the evidence that the performance degradation is observed mainly on dual episodes (Table 11). While many clinicians are able to perform equally well on their dual episodes and non-dual episodes, there are many more clinicians who are performing significantly worse on their dual episodes than their non-dual episodes, which suggests that clinicians aren't able to fully mitigate the effect of SRFs (Table 12). Lastly, risk adjusting for dual status appears to change the performance ranking for many clinicians (Table 13).

Table 10: Coefficient of Patient-level Dual Status under Different Models

Level	Subgroup Risk Model	% of All Episodes	Coefficient of Patient-level Dual Status in Log-transformed Episode Cost (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 Dual Status + Fixed Effect
TIN	Single Subgroup with Part D Coverage	77.09%	0.29 (SD: 0.01, p: <0.001)	0.4 (SD: 0.01, p: <0.001)	0.4 (SD: 0.01, p: <0.001)
TIN	Single Subgroup without Part D Coverage	22.91%	0.04 (SD: 0.07, p: 0.54)	0.08 (SD: 0.07, p: 0.25)	0.1 (SD: 0.07, p: 0.18)
TIN-NPI	Single Subgroup with Part D Coverage	77.31%	0.33 (SD: 0.01, p: <0.001)	0.5 (SD: 0.01, p: <0.001)	0.49 (SD: 0.01, p: <0.001)
TIN-NPI	Single Subgroup without Part D Coverage	22.69%	0.03 (SD: 0.09, p: 0.69)	0.12 (SD: 0.09, p: 0.16)	0.08 (SD: 0.1, p: 0.41)

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

Dual Share	TIN			TIN-NPI		
	All Episodes	Dual Episodes	Non-Dual Episodes	All Episodes	Dual Episodes	Non-Dual Episodes
All	1.00	1.23	0.95	1.03	1.31	0.98
0%	0.91		0.91	0.99		0.99
1-20%	1.00	1.24	0.97	1.03	1.30	1.00
21-40%	0.99	1.20	0.91	1.03	1.35	0.91
41-60%	1.01	1.09	0.92	1.12	1.34	0.92
61-80%	1.10	1.20	0.87	1.14	1.33	0.73
81-99%	1.22	1.27	0.88	0.84	0.89	0.53
100%	0.77	0.77		0.39	0.39	

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

Reporting Level	Significantly Worse	Equally Well	Significantly Better
TIN	17%	82%	1%
TIN-NPI	16%	83%	1%

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

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	79.1%	12.9%
TIN-NPI	80.2%	13.1%

2.8 Impact of Exclusions

Table 14 displays descriptive statistics of all episodes meeting the 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 remains relatively similar after applying the exclusions, with changes from the mean of \$11,990 to \$11,708 at the TIN level and \$12,733 at the TIN-NPI level (Table 14). Most of the exclusion criteria have higher mean observed cost than all episodes meeting triggering logic, with the exception of TINs or TIN-NPIs not meeting the case minimum or there not being an attributed NPI.

Episodes shorter than one year are excluded because the methodology for the chronic measures requires at least one year of claims data to measure clinician cost performance to ensure sufficient observation of chronic care, which is often intermittent and sparse over a long period of time. Although these episodes are excluded during the performance period being examined, they are likely to be included in the following performance period once the episode length is longer than one year.

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 \$14,954 (Table 14), likely because the costs are distributed over fewer days than a typical episode.

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 \$21,625 compared to \$11,990 for all episodes meeting triggering logic (Table 14). The wide variability of observed episode costs for outlier cases also supports their exclusion. At the 10th percentile the outlier cases observed cost is \$72 and at the 90th percentile the observed cost is \$52,676.

Episodes where there is not an attributed clinician are excluded because these episode do not have any TIN-NPIs that billed at least 30% of the clinically-related claims with a relevant diagnosis. As such, they cannot be used in the measure at the TIN-NPI level.

Table 14: Cost Statistics for Measure Exclusions

Exclusion Criteria	Episodes		Mean	Observed Episode Cost Percentile				
	Count	Percent of All Episodes Meeting Trigger Logic		10 th	25 th	50 th	75 th	90 th
All Episodes Meeting Triggering Logic	574,258	100.00%	\$11,990	\$389	\$807	\$2,297	\$18,896	\$40,004
Episode Length Less Than One Attribution Window	12,190	2.12%	\$15,324	\$732	\$1,639	\$5,830	\$17,762	\$38,776
Beneficiary Death in Episode	41,158	7.17%	\$14,954	\$606	\$1,492	\$6,333	\$20,641	\$40,825
Outlier	10,660	1.86%	\$21,625	\$72	\$98	\$5,245	\$45,325	\$52,676
No Attributed NPI (TIN-NPI reporting only)	205,286	35.75%	\$11,259	\$321	\$759	\$2,227	\$16,596	\$38,572
TIN does not Meet Case Minimum	37,998	6.62%	\$9,976	\$287	\$665	\$1,731	\$12,764	\$36,710
TIN-NPI does not Meet Case Minimum	93,961	16.36%	\$10,255	\$313	\$693	\$1,749	\$13,630	\$37,441
Reportable Episodes (if all clinicians reported as TIN at the Testing Volume Threshold)	488,906	85.14%	\$11,708	\$409	\$808	\$2,143	\$18,637	\$39,446
Reportable Episodes (if all clinicians reported as TIN-NPI at the Testing Volume Threshold)	255,657	44.52%	\$12,733	\$488	\$870	\$2,487	\$21,478	\$40,748

Appendix A. Distributions of Measure Score

Figure 2: Distribution of Measure Score - TIN

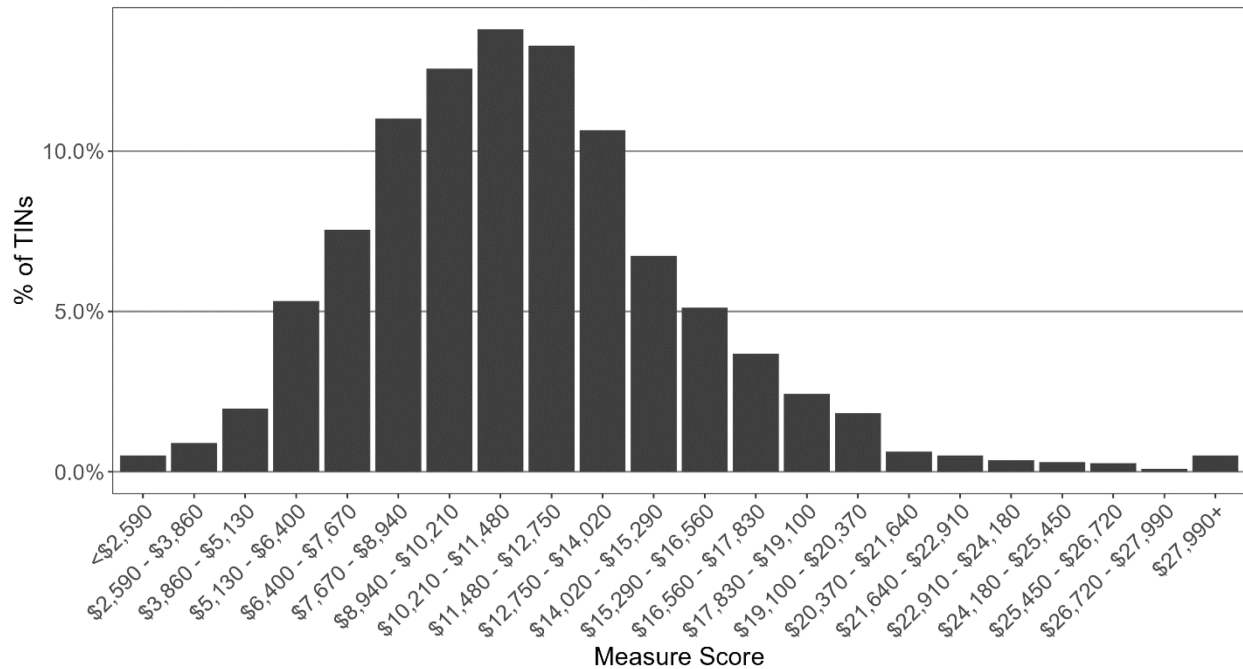


Figure 3: Distribution of Measure Score - TIN-NPI

