

# **Parkinsonism Syndromes and Multiple Sclerosis (MS)**

## **Measure Testing Form**

January 2026



# Contents

<b>1.0</b>	<b>Introduction.....</b>	<b>3</b>
1.1	Project Title and Overview.....	3
1.2	Measure Name.....	3
1.3	Type of Measure .....	3
1.4	Data.....	3
<b>2.0</b>	<b>Preliminary Testing Results .....</b>	<b>4</b>
2.1	Measure Coverage.....	4
2.2	Frequently Attributed Specialties .....	5
2.3	Reliability .....	5
2.4	Validity .....	6
	2.4.1 Conceptual Model.....	6
	2.4.2 Empirical Validity Analyses.....	7
2.5	Performance Gap.....	10
2.6	Risk Adjustment and Stratification .....	11
	2.6.1 Discrimination.....	11
	2.6.2 Calibration.....	12
2.7	Dual Status and Provider Location Analysis.....	12
2.8	Impact of Exclusions .....	15
	<b>Appendix A. Glossary .....</b>	<b>18</b>
	<b>Appendix B. Distributions of Measure Score .....</b>	<b>20</b>

# 1.0 Introduction

This Measure Testing Form (MTF) provides a summary of the preliminary measure testing results as part of field testing the Parkinsonism Syndromes and MS episode-based cost measure. Field testing, also known as beta testing, is part of the measure development and testing process outlined by the Centers for Medicare & Medicaid Services (CMS) Measures Management System (MMS).<sup>1</sup> This testing occurs after development of initial technical specifications and is an opportunity to assess scientific acceptability (e.g., the extent to which the measure produces valid and reliable results about the intended area of measurement) and usability of a measure (e.g., whether people or organizations can effectively use a measure). Field testing also provides additional evidence to support importance and feasibility evaluations. More information about measure evaluation criteria is available on the MMS Hub.<sup>2</sup>

Readers may review aggregate field testing results in this document, alongside other measure documentation and field testing reports, 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.<sup>3</sup>

## 1.1 Project Title and Overview

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 75FCMC24F0142.

## 1.2 Measure Name

Parkinsonism Syndromes and MS Episode-Based Cost Measure

## 1.3 Type of Measure

Cost/Resource Use

## 1.4 Data

The study period is from January 1, 2024, through December 31, 2024. 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, the Long-Term Minimum Data Set, and the Medicare Enrollment Database.

Testing results are presented at a testing volume threshold of 20 episodes for clinician groups and individual clinicians. Throughout this document, 'Group' refers to clinician groups identified by a Tax Identification Number (TIN), and 'Clinician' refers to individual clinicians identified by a combination of a TIN and National Provider Identifier (TIN-NPI).

---

<sup>1</sup> MMS, “Measure Testing Process,” <https://mmshub.cms.gov/measure-lifecycle/measure-testing/process/overview>

<sup>2</sup> MMS, “Measure Evaluation Criteria,” <https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria>

<sup>3</sup> These documents will be available on the CMS Cost Measures Information page once field testing begins: <https://www.cms.gov/medicare/quality/value-based-programs/cost-measures/current>.

## 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. Section 2.2 lists the most frequently attributed specialties. Section 2.3 presents reliability testing results. Section 2.4 includes conceptual and empirical validity information. Section 2.5 demonstrates the measure's performance gap. Section 2.6 presents empirical results of the risk adjustment and stratification methods used by this measure. Section 2.7 examines the associations between non-clinical factors—including dual Medicare and Medicaid enrollment status and provider location (urban versus rural)—and measure performance. Section 2.8 examines the impact of exclusion criteria used by the measure through their frequency and resource use patterns. Appendix A provides a glossary of key terms and definitions.

### 2.1 Measure Coverage

Table 1 shows the patient population for the Parkinsonism Syndromes and MS measure testing. It consists of Medicare beneficiaries enrolled in Medicare Parts A and B who meet all inclusion and exclusion criteria as specified by the measure.

**Table 1: Beneficiary Demographics**

Metric	Value
Number of Beneficiaries	294,337
Mean Age	74.2
Part D Enrollment %	80.31%

Table 2 shows the characteristics of groups and clinicians who are attributed at least 20 episodes.

**Table 2: Group and Clinician Characteristics**

Metric	Group		Clinician	
	Count	%	Count	%
Count	3,115	100.00%	2,884	100.00%
<b>Number of Episodes Attributed</b>	-	-	-	-
20-39 Episodes	1,318	42.31%	1,735	60.16%
40-59 Episodes	561	18.01%	561	19.45%
60-79 Episodes	249	7.99%	242	8.39%
80-99 Episodes	175	5.62%	135	4.68%
100-199 Episodes	433	13.90%	186	6.45%
200-299 Episodes	148	4.75%	23	0.80%
300+ Episodes	231	7.42%	2	0.07%
<b>Census Region</b>	-	-	-	-
Northeast	619	19.87%	594	20.60%
Midwest	627	20.13%	557	19.31%
South	1,235	39.65%	1,095	37.97%
West	632	20.29%	636	22.05%
Unknown	2	0.06%	2	0.07%

## 2.2 Frequently Attributed Specialties

Table 3 shows the top 10 attributed specialties for this measure determined by the official CMS specialty designations listed on Medicare claims, 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 Parkinsonism Syndromes and MS, with neurology being the most commonly attributed specialty. These clinicians are also consistent with input provided by stakeholders, including patient and family partners (PFPs), during the measure development process. PFPs identified a similar range of clinicians involved in their care, such as family practitioners, nurse practitioners, neurologists, physical therapists, physician assistants, and cardiologists.

**Table 3: Count of the Top 10 Attributed Specialties**

Specialty	Number of Clinicians Attributed
Neurology	2,380
Nurse Practitioner	248
Physician Assistant	107
Internal Medicine	49
Physical Medicine and Rehabilitation	27
Family Practice	16
Psychiatry	16
Neuropsychiatry	13
Nephrology	5
Certified Clinical Nurse Specialist	5

## 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 testing 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 Groups and Clinicians above Moderate Reliability at Various Testing Volume Thresholds**

Testing Volume Threshold	Group			Clinician		
	Number of Groups	Mean Reliability	Percent Above 0.4	Number of Clinicians	Mean Reliability	Percent Above 0.4
10	5,544	0.55	71.19%	6,171	0.51	64.98%
<b>20</b>	<b>3,115</b>	<b>0.66</b>	<b>89.05%</b>	<b>2,884</b>	<b>0.60</b>	<b>81.93%</b>
30	2,283	0.71	94.09%	1,753	0.65	89.28%

At the testing volume of 20 episodes, the mean reliability for the Parkinsonism Syndromes and MS measure is moderate, specifically 0.66 at the group level and 0.60 at the clinician 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.<sup>4,5</sup> Most groups and clinicians meet or exceed the moderate reliability threshold of 0.4 at the 20-episode testing volume threshold.

## 2.4 Validity

Validity is a criterion that evaluates whether the cost measure is able to quantify the construct that it aims to measure. Cost measures are designed to reflect the cost directly related to clinicians’ treatment choices, as well as the cost of adverse outcomes as a result of care (e.g., inpatient hospitalizations, emergency room visits, and post-acute care that occur after the episode starts).

Cost measure validity can be demonstrated in a conceptual model (Section 2.4.1). Validity can also be evaluated empirically by examining whether measure performance aligns with expectations from the conceptual model. That is, analyses can show to what extent adverse outcomes and treatment choices contribute to measure performance (Section 2.4.2).

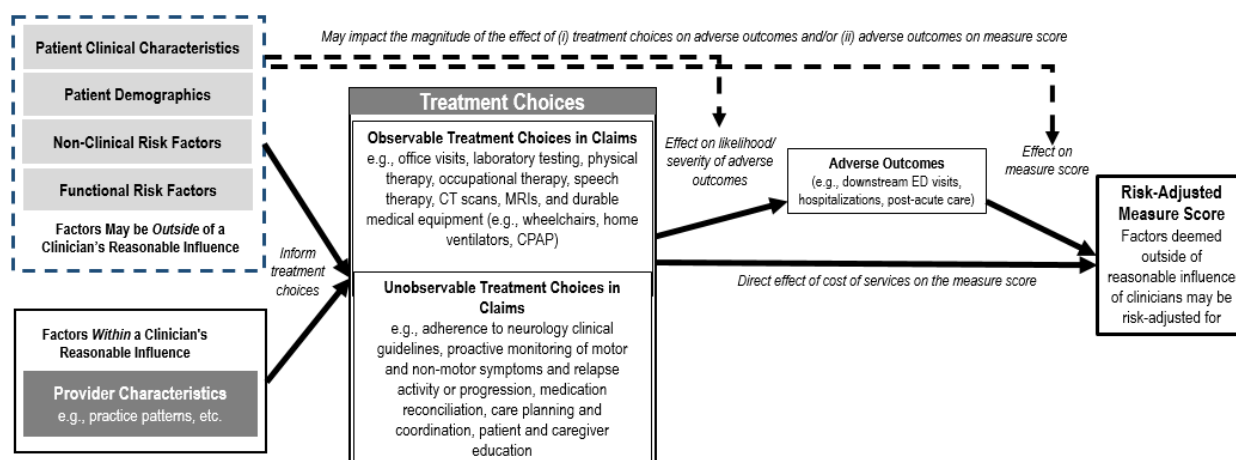
### 2.4.1 Conceptual Model

The measure conceptual model creates a comprehensive picture of the factors in and outside of a clinician’s control and how these factors interact with treatment choices, adverse outcomes, and the final measure score. Factors outside of a clinician’s reasonable influence include patient characteristics, patient demographics, and preexisting risk factors. These factors, combined with provider characteristics, impact the type, magnitude, and severity of treatment choices and adverse care outcomes. The conceptual model shown in Figure 1 illustrates this relationship as well as how treatment choices and adverse outcomes impact the cost of services and affect the measure score.

<sup>4</sup> 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](http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability-.pdf).

<sup>5</sup> 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/2021/01/26/2021-02434).

**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. For the Parkinsonism Syndromes and MS cost measure, clinically relevant treatment choices such as physical therapy, occupational therapy, speech therapy,, laboratory testing, imaging services (e.g., CT scans and MRIs), and the use of durable medical equipment (e.g., wheelchairs, home ventilators, and CPAP) all factor into the potential adverse outcomes associated with each episode. Adverse outcomes for this cost measure are typically emergency services, hospitalizations/inpatient stays, and the use of post-acute care such as home health, skilled nursing facilities, and long term rehabilitative care.

## 2.4.2 Empirical Validity Analyses

To empirically assess validity, we first examined the impact of adverse events on episode performance, before and after risk adjustment (Table 5). Next, we assessed the impact of treatment choices on occurrence of adverse events and measure performance (Table 6).

Table 5 presents the distribution of observed and risk-adjusted costs for episodes with adverse outcomes compared to all episodes and episodes with outpatient evaluation and management (E/M) services. Outpatient E/M services are included as a reference point, as these services occur in nearly all episodes and represent routine care delivery.

**Table 5: Adverse Events Effect on Episode Cost**

Categories of Service	% of Episodes	Observed Cost		Risk-Adjusted Cost	
		Mean	Std. Dev.	Mean	Std. Dev.
All Episodes	100.00%	\$13,451	\$24,131	\$14,017	\$20,182
Outpatient E/M Services	99.99%	\$13,451	\$24,130	\$14,017	\$20,183
Adverse Outcomes	-	-	-	-	-
Emergency E/M Services	28.07%	\$21,040	\$30,965	\$21,053	\$22,844
Inpatient Hospital	12.18%	\$37,555	\$35,783	\$37,788	\$29,289
Skilled Nursing Facility	10.86%	\$28,658	\$28,680	\$27,791	\$28,188
Inpatient Rehabilitation or Long-Term Care Hospital	1.33%	\$57,002	\$34,535	\$56,864	\$35,551

Table 5 demonstrates that Parkinsonism Syndromes and MS episodes with adverse outcomes are substantially more costly than episodes overall. Episodes involving inpatient rehabilitation or long-term care hospital stays have a mean observed cost of \$57,002, more than four times higher than the mean observed cost of \$13,451 for all episodes. Similarly, episodes with inpatient hospitalizations (\$37,555), skilled nursing facility stays (\$28,658), and emergency department visits (\$21,040) all have mean observed costs that are 1.5 to 3 times higher than the baseline.

After risk adjustment, episodes with adverse outcomes continue to show higher costs than all episodes, with minimal attenuation. For example, the mean risk-adjusted cost for episodes with inpatient rehabilitation or long-term care hospital stays is \$56,864, compared to \$14,017 for all episodes. The persistence of higher costs after risk adjustment, particularly for inpatient rehabilitation, inpatient hospitalizations, and skilled nursing facility stays, demonstrates that these adverse outcomes meaningfully contribute to measure performance beyond what is explained by patient-level risk factors.

These results support the validity of the Parkinsonism Syndromes and MS cost measure by demonstrating that adverse outcomes, which represent potentially avoidable complications or exacerbations, contribute significantly to measure performance. Clinicians whose treatment choices reduce the likelihood of hospitalizations, emergency visits, and post-acute care utilization would be expected to have lower measure scores, consistent with the measure's intent to incentivize high-value care for patients with progressive neurological conditions. Table 6 presents results from two regression models designed to evaluate whether the measure score reflects both direct and indirect effects of treatment choices:

- Model 1 (Direct Effect): Estimates the association between treatment choices and the measure score while controlling for the cost of adverse outcomes.
- Model 2 (Indirect Effect): Estimates the association between treatment choices and the cost of adverse outcomes.

Both models use the mean cost across episodes attributed to each clinician. The measure score is represented by a group or clinician's mean observed-to-expected cost ratio (O/E ratio)—the ratio of actual costs to costs predicted by risk adjustment—across attributed episodes.

Table 6 tests whether treatment choices affect measure performance as expected based on the conceptual model. A positive coefficient indicates that higher use of a service category is associated with higher costs; a negative coefficient indicates lower costs. The 95% confidence interval shows the range of plausible values for each estimate, and the p-value indicates statistical significance (values below 0.05 suggest the finding is unlikely to be due to chance).

- Model 1 answers: "Do treatment choices affect the measure score directly, after accounting for adverse events, demonstrating direct effects?"
- Model 2 answers: "Do treatment choices affect the cost of adverse events, demonstrating indirect effects?"

**Table 6: Estimated Model Effect of Adverse Events and Treatment Choices**

Categories of Service	Coefficient in Thousands of Dollars [95% Confidence Interval] (p-value)			
	Group		Clinician	
	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	0.05 [0.05,0.05] (p < 0.01)	-	0.06 [0.05,0.06] (p < 0.01)	-
Ambulatory/Minor Procedures	0.10 [0.01,0.18] (p = 0.03)	-0.70 [-1.67,0.27] (p = 0.16)	0.12 [0.04,0.20] (p < 0.01)	-0.50 [-1.13,0.14] (p = 0.13)
Part B Drugs	0.03 [0.03,0.03] (p < 0.01)	-0.04 [-0.07,-0.01] (p = 0.02)	0.03 [0.03,0.03] (p < 0.01)	-0.01 [-0.03,0.01] (p = 0.23)
Part D Drugs	0.03 [0.02,0.03] (p < 0.01)	0.03 [-0.01,0.08] (p = 0.15)	0.02 [0.02,0.03] (p < 0.01)	0.02 [-0.01,0.04] (p = 0.22)
Durable Medical Equipment and Supplies	0.00 [-0.01,0.00] (p = 0.39)	0.23 [0.17,0.28] (p < 0.01)	0.00 [-0.01,0.01] (p = 0.66)	0.19 [0.11,0.27] (p < 0.01)
Imaging Services	0.23 [0.09,0.36] (p < 0.01)	-6.35 [-7.81,-4.90] (p < 0.01)	-0.05 [-0.16,0.05] (p = 0.31)	-1.42 [-2.24,-0.61] (p < 0.01)
Laboratory, Pathology, and Other Tests	0.17 [0.04,0.30] (p = 0.01)	-4.36 [-5.79,-2.94] (p < 0.01)	0.43 [0.30,0.55] (p < 0.01)	-2.85 [-3.82,-1.88] (p < 0.01)
Major Procedures	0.33 [0.22,0.44] (p < 0.01)	-1.81 [-3.00,-0.61] (p < 0.01)	0.20 [0.11,0.28] (p < 0.01)	-1.01 [-1.70,-0.32] (p < 0.01)
Outpatient Evaluation & Management Services	-0.10 [-0.12,-0.08] (p < 0.01)	4.59 [4.40,4.79] (p < 0.01)	-0.08 [-0.12,-0.05] (p < 0.01)	3.70 [3.45,3.95] (p < 0.01)
Outpatient Physical, Occupational, or Speech and Language Pathology Therapy	0.07 [0.06,0.09] (p < 0.01)	-0.30 [-0.47,-0.13] (p < 0.01)	0.09 [0.08,0.11] (p < 0.01)	-0.24 [-0.38,-0.10] (p < 0.01)

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 6).

Model 1 demonstrates many adverse events and treatment service categories have a statistically significant associations with the measure score. There is a statistically significant positive association between costs of adverse events and measure score at the clinician- and group-level, meaning providers with higher adverse event costs have higher episode costs

(worse measure performance). Additionally, there is a statistically significant positive association between measure scores and costs of ambulatory/minor procedures; Part B Drugs; Part D Drugs; imaging services (group only); laboratory, pathology, and other tests; major procedures; and outpatient physical, occupational, or speech and language pathology services. These service categories represent discretionary treatment choices that can directly increase episode costs. However, outpatient evaluation & management (E/M) services have a statistically significant negative association with measure score, suggesting that higher utilization of E/Ms could improve cost performance (e.g., through increased care coordination and management).

Model 2 shows that treatment choices can influence costs of adverse events, thereby affecting cost performance. For example, higher E/M and DME utilization is associated with higher adverse event costs. However, this should be interpreted in context: patients requiring more frequent outpatient visits or DME may have greater disease severity or complexity that predisposes them to adverse events. Alternatively, patients may require increased frequency of E/M visits or use of DME after an adverse event. Several treatment choices demonstrate significant negative indirect effects, meaning higher utilization is associated with lower adverse event costs, including imaging services and laboratory/pathology tests, suggesting that appropriate diagnostic monitoring may help prevent costly adverse outcomes. Similarly, major procedures and outpatient therapy services are associated with reduced adverse event costs, consistent with the clinical expectation that timely procedural intervention and rehabilitation services can mitigate complications in this patient population and potentially improve cost performance.

## 2.5 Performance Gap

Table 7 shows the distribution of the measure score for clinicians and clinician groups meeting the testing volume threshold. These results align with expectations based on our review of the literature and demonstrate that there is a performance gap in cost measure performance between the most and least efficient groups and clinicians. Appendix B includes histograms for the distribution of group and clinician scores.

**Table 7: Distribution of the Measure Score**

Metric	Group	Clinician
Mean Score	\$14,064	\$13,944
Score Interquartile Range (IQR)	\$4,474	\$5,745
Standard Deviation	\$4,095	\$4,991
Coefficient of Variation	0.29	0.36
Score Percentile	Group	Clinician
10 <sup>th</sup>	\$9,416	\$8,528
25 <sup>th</sup>	\$11,529	\$10,623
50 <sup>th</sup>	\$13,685	\$13,268
75 <sup>th</sup>	\$16,002	\$16,368
90 <sup>th</sup>	\$18,850	\$20,015

The Parkinsonism Syndromes and MS cost measure score distribution shows variation in performance across clinicians and groups. For example, the 90<sup>th</sup> percentile measure score is more than 2 times greater than the 10<sup>th</sup> percentile measure score for both the groups and clinicians. The variation in the measure score, indicated by the interquartile range and standard deviation, ranges from \$4,000 to nearly \$6,000, depending on the metric and reporting level.

based on the reporting level. The results suggest that there is an opportunity for the least efficient clinicians and groups to improve their cost performance.

## 2.6 Risk Adjustment and Stratification

As shown in the conceptual model (Figure 1), patient-level and clinician-level factors can influence the measure score, which is informed by both published external research and our own data analysis.<sup>6,7,8,9</sup> 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 non-clinical factors, 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 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 adjusters 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 eligibility for dual Medicare and Medicaid enrollment 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-squared value for the measure is 0.202, and 0.201 after adjusting for the model's complexity based on the number of risk adjusters used. In other words, 20% 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 adjusters and sub-groups.

---

<sup>6</sup>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.

<sup>7</sup>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

<sup>8</sup>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/>.

<sup>9</sup>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/social-risk-factors-and-medicare-value-based-purchasing-programs>

## 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. Table 8 assesses calibration by calculating the average predictive ratio for episodes partitioned into deciles based on expected episode cost. The predictive ratio is calculated using the formula of average expected cost divided by 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 8: Predictive Ratio by Decile of Predicted Episode Cost**

Decile	Average Predictive Ratio
All	1.00
Decile 1	0.80
Decile 2	0.86
Decile 3	0.97
Decile 4	0.96
Decile 5	0.95
Decile 6	0.93
Decile 7	0.89
Decile 8	0.89
Decile 9	0.97
Decile 10	1.20

Table 8 shows that the model has an overall predictive ratio of 1.0, with only low or moderate variation in predictive ratios across most risk score deciles. However, the lowest and highest risk score deciles suggest possible opportunities to improve risk classification for the lowest and highest risk patients, with Decile 1 showing under- prediction (0.80) and Decile 10 showing over-prediction (1.20).

## 2.7 Dual Status and Provider Location Analysis

Beyond clinical characteristics of patients, the cost of care may be influenced by non-clinical factors such as location or coverage eligibility. Beneficiaries located in rural versus urban areas may face different barriers to accessing care. Additionally, beneficiaries eligible for dual Medicare and Medicaid enrollment status have been historically shown as a more vulnerable population, more likely to experience poor outcomes. At the program level, MIPS adjusts for these factors using the MIPS Complex Patient Bonus to ensure clinicians or groups treating more complex patients are not disadvantaged.<sup>10</sup> At the measure-level, 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.

---

<sup>10</sup> Quality Payment Program “COVID-19 Response,” <https://qpp-cm-prod-content.s3.amazonaws.com/uploads/966/QPP%20COVID-19%20Response%20Fact%20Sheet.pdf>

Table 9 outlines the advantages and disadvantages of using dual status and provider claim location as indicators of non-clinical care factors.

**Table 9: Tradeoffs of Dual Status and Claim Location Indicators**

Variable	Advantages	Disadvantages	Used in Testing
Dual Medicare and Medicaid enrollment status	<ul style="list-style-type: none"> <li>Available for all beneficiaries</li> <li>Most powerful predictor of poor outcomes<sup>11</sup></li> </ul>	<ul style="list-style-type: none"> <li>Variation in Medicaid eligibility across states</li> </ul>	Yes
Provider Location (i.e., urban or rural)	<ul style="list-style-type: none"> <li>Urban or rural provider location can reflect potential differences in patient access or barriers to care</li> </ul>	<ul style="list-style-type: none"> <li>Reflects provider location, not beneficiary residence, which may differ</li> </ul>	Yes

Table 10 presents the distribution of measure performance for groups and clinicians by provider location. Provider location (urban or rural) is determined by the Part B claim billing zip code capturing the highest total annual standardized cost from the 2024 study period or any year within a 2 year lookback window. The selected zip code per provider is then mapped to its zip locality 'rural' designation using the most recent quarter of mapping data available for the 2024. Provider scores close to the national average observed episode cost for both urban and rural providers would indicate that the risk adjustment model adequately accounts for geographic differences in cost patterns.

**Table 10: Associations of Provider Location for Groups and Clinicians**

Provider Level	Provider Location	% of All Providers	Mean Provider Score	Distribution of Provider Score (Percentiles)				
				P10	P25	P50	P75	P90
Group	All Providers	100.00%	\$14,064	\$9,416	\$11,529	\$13,685	\$16,002	\$18,850
	Urban	87.22%	\$14,162	\$9,588	\$11,697	\$13,782	\$16,037	\$18,931
	Rural	12.78%	\$13,395	\$8,428	\$10,605	\$12,886	\$15,518	\$18,461
Clinician	All Providers	100.00%	\$13,944	\$8,528	\$10,623	\$13,268	\$16,368	\$20,015
	Urban	92.09%	\$14,028	\$8,629	\$10,718	\$13,325	\$16,406	\$20,031
	Rural	7.91%	\$12,964	\$7,188	\$9,333	\$12,436	\$15,935	\$19,904

The majority of clinicians and groups practice in an urban location. Results show that clinicians and groups in rural locations perform similarly to those in urban locations, with slightly better performance observed for clinicians in rural locations. These findings suggest that additional adjustment for geographic location is not needed.

<sup>11</sup> Refer to footnote 4.

The subsequent analyses focus on dual status as the main proxy indicator of non-clinical factors for risk adjustment. Dual status indicates that a beneficiary is simultaneously enrolled in both Medicare and Medicaid. For testing purposes, a beneficiary will only be flagged as having dual status if they are indicated as such in the Common Medicare Environment throughout the entirety of the episode window. To determine whether it's appropriate to risk adjust for dual status eligibility, the following criteria are considered:

- Association (Table 11): Is dual status associated with measure performance?
- Source (Table 11): Is the association patient-level or clinician-level?
- Complexity versus Quality (Tables 12, 13): Is the patient need, not poor quality, driving differences?
- Impact (Table 14): How would dual status adjustment affect measure performance rankings?

**Table 11: 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 (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
Group	Parkinson's and Related Conditions without Part-D Enrollment	15.46%	0.19 (p < 0.0001)	0.10 (p = 0.04)	0.14 (p = 0.02)
	Parkinson's and Related Conditions with Part-D Enrollment	61.08%	0.14 (p < 0.0001)	0.10 (p < 0.0001)	0.10 (p < 0.0001)
	Multiple Sclerosis without Part-D Enrollment	4.33%	0.08 (p = 0.32)	0.03 (p = 0.68)	0.02 (p = 0.83)
	Multiple Sclerosis with Part-D Enrollment	19.14%	0.43 (p < 0.0001)	0.41 (p < 0.0001)	0.40 (p < 0.0001)
Clinician	Parkinson's and Related Conditions without Part-D Enrollment	15.38%	0.18 (p = 0.0004)	0.09 (p = 0.09)	0.16 (p = 0.06)
	Parkinson's and Related Conditions with Part-D Enrollment	61.37%	0.14 (p < 0.0001)	0.15 (p < 0.0001)	0.13 (p < 0.0001)
	Multiple Sclerosis without Part-D Enrollment	4.28%	0.09 (p = 0.29)	-0.05 (p = 0.60)	0.04 (p = 0.82)
	Multiple Sclerosis with Part-D Enrollment	18.97%	0.45 (p < 0.0001)	0.45 (p < 0.0001)	0.48 (p < 0.0001)

**Table 12: Mean O/E Ratio Stratified by Dual Share and Patient's Dual Status**

Share of Episodes with Dual Status	Group			Clinician		
	All Episodes	Dual Episodes	Non-Dual Episodes	All Episodes	Dual Episodes	Non-Dual Episodes
All	1.08	1.16	1.07	1.08	1.26	1.07
0-20%	1.03	1.19	1.03	1.03	1.36	1.03
21-40%	1.05	1.19	1.04	1.04	1.30	1.02
41-60%	1.09	1.18	1.08	1.06	1.24	1.05
61-80%	1.10	1.14	1.09	1.10	1.22	1.08
81-100%	1.13	1.11	1.13	1.18	1.25	1.18

**Table 13: Proportions of Groups and 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
Group	7.22%	90.63%	2.15%
Clinician	9.00%	90.30%	0.70%

**Table 14: Group and Clinician Performance Shift Measured by the Change in Measure Score After Adding a Patient-level Dual Status Risk Adjustor**

Reporting Level	Proportions of Groups and Clinicians Affected at Various Levels of Performance Shift	
	Ranking Shift by 1% or more	Ranking Shift by 5% or more
Group	74.54%	10.98%
Clinician	63.52%	8.78%

Among subgroups with Part D enrollment, which comprise the majority of episodes, there is a significant association between a patient's dual status and episode cost, as observed in Table 11. Very little shift in model coefficients is seen when adding in clinician-level factors, suggesting that patient-level dual status has a greater effect on the model than any clinician-level factors. Additionally, patients with dual status tend to have higher mean ratios of observed to expected costs when compared to all episodes and non-dual episodes (Table 12). When looking at group and clinician performance for dual versus non-dual episodes, approximately 90% of groups and clinicians perform equally well for dual and non-dual episodes (Table 13). However, fewer groups and clinician perform significantly better for dual episodes than those who perform significantly worse for dual episodes. Further, approximately 10 percent of clinicians and groups would see a ranking shift of more than 5 percent after adjusting for dual status (Table 14). These results suggest that it may be appropriate to continue to adjust for dual status.

## 2.8 Impact of Exclusions

Table 15 displays descriptive statistics of all episodes meeting the measure's triggering logic, excluded episodes, and final reportable episodes at both group and clinician 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.

**Table 15: Cost Statistics for Measure Exclusions**

Exclusion Criteria	Episodes		Observed Episode Cost					
	Count	Percent of All Episodes Meeting Trigger Logic	Mean	Percentile				
				10 <sup>th</sup>	25 <sup>th</sup>	50 <sup>th</sup>	75 <sup>th</sup>	90 <sup>th</sup>
All Episodes Meeting Triggering Logic	510,170	100.00%	\$15,752	\$687	\$1,621	\$5,630	\$18,461	\$46,401
Episode Length Less Than 1 Year	29,669	5.82%	\$37,346	\$1,690	\$4,762	\$15,098	\$41,340	\$92,427
Beneficiary Death in Episode	70,145	13.75%	\$28,994	\$1,511	\$4,607	\$14,346	\$34,572	\$67,701
Outlier Cases	8,749	1.71%	\$42,747	\$2,488	\$7,717	\$55,540	\$63,524	\$94,250
No Attributed Clinician (Clinician Reporting Only)	47,402	9.29%	\$18,122	\$1,018	\$2,461	\$7,693	\$22,858	\$51,874
Group does not Meet Testing Volume Threshold	121,309	23.78%	\$17,291	\$670	\$1,723	\$6,414	\$20,701	\$49,330
Clinician does not Meet Testing Volume Threshold	313,856	61.52%	\$16,262	\$647	\$1,592	\$5,847	\$19,236	\$47,318
Microvascular Decompression <sup>12</sup>	-	-	-	-	-	-	-	-
Spinal Cord Injury	32	0.01%	\$35,387	\$1,460	\$2,757	\$18,676	\$32,016	\$50,647
Stereotactic Radiosurgery	75	0.01%	\$26,942	\$1,229	\$2,625	\$9,327	\$41,863	\$75,676
<b>Reportable Episodes</b> (if all clinicians reported as Group at the Testing Volume Threshold)	330,023	64.69%	\$12,746	\$641	\$1,407	\$4,518	\$14,497	\$38,033
<b>Reportable Episodes</b> (if all clinicians reported as Clinician at the Testing Volume Threshold)	132,357	25.94%	\$12,005	\$661	\$1,367	\$4,085	\$12,875	\$35,832

Overall, exclusion criteria reduced episode volume from 510,170 episodes to a final population of 330,023 group-level and 132,357 clinician-level episodes. The reduction in episodes is primarily driven by standard exclusion criteria, with the largest exclusion categories being episodes for groups and clinicians who did not meet the testing volume threshold. Other standard exclusions are intended to ensure sufficient comparable data for assessment (i.e.,

<sup>12</sup> Exclusions with less than 11 reportable cases are suppressed from publication.

excluding episode less than one year, episodes ending in death, and outlier cases. Very few episodes were excluded due to measure-specific exclusions (i.e., episodes with prior microvascular decompression, spinal cord injury, or stereotactic radiosurgery).

Generally, mean observed costs for excluded patient cohorts were higher than mean observed cost for the full episode population as well as reportable episodes for groups and clinicians. The differences in cost distribution for these episodes support their continued exclusion from the measure.

Collectively, these findings support the continues exclusion of these episodes to ensure a comparable patient cohort that will yield a clinically coherent measure and meaningful information to attributed groups and clinicians.

## Appendix A. Glossary

**Table A1: Key Terms and Definitions**

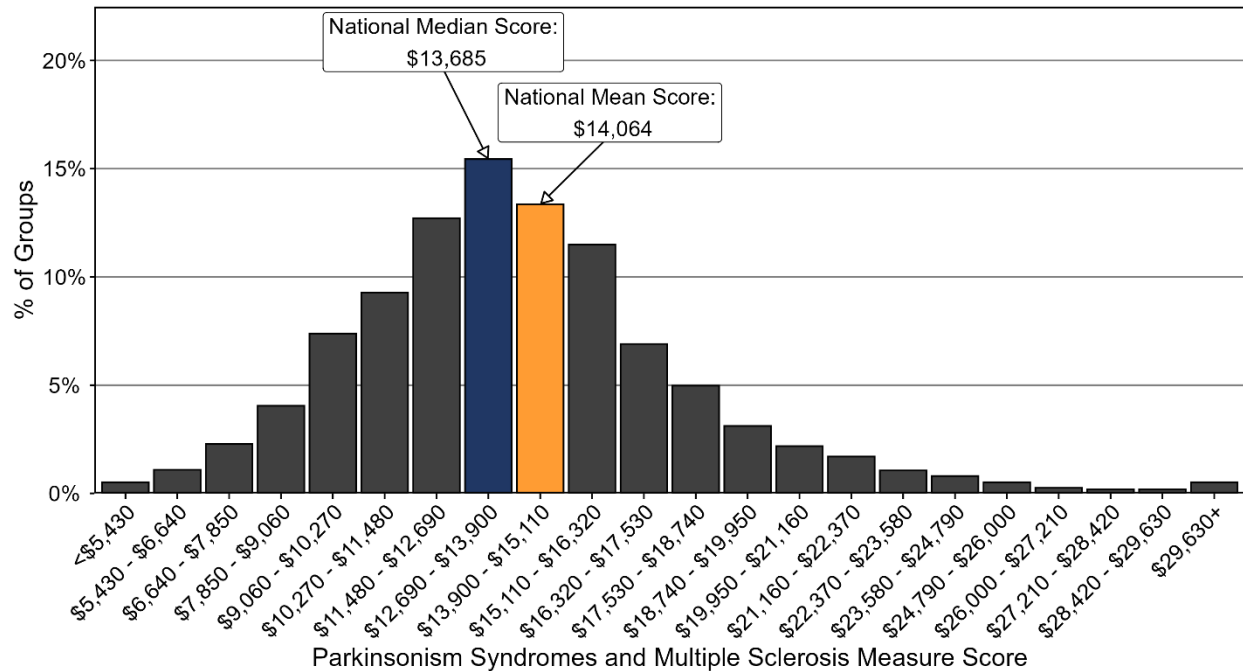
Term	Definition
Attributed Clinician/ Group	The clinician or group determined to be responsible for an episode based on the measure's attribution methodology. Please see the Draft Cost Measure Methodology for a description of the attribution methodology and the Draft Measure Codes List for the list of codes in measure attribution. <sup>13</sup>
Calibration	A statistical property that evaluates whether a risk adjustment model produces accurate cost predictions across the full range of patient complexity. A well-calibrated model has predictive ratios close to 1.0 across all risk levels.
Coefficient of Variation	A measure of relative variability, calculated as the standard deviation divided by the mean. Higher values indicate greater variability in measure scores.
Discrimination	A statistical property that evaluates how well a risk adjustment model distinguishes between high-cost and low-cost episodes. Measured by R-squared.
Dual Eligible (Dual Status)	Beneficiaries who qualify for both Medicare and Medicaid. Dual-eligible beneficiaries often have greater health needs and may face additional barriers to care.
Episode	A defined period during which clinically related services are delivered to a patient for a specific condition or procedure. Episode-based cost measures group these services together to assess resource use.
Expected Cost	The cost predicted by the risk adjustment model for an episode, based on patient characteristics, risk factors, and other variables included in the model. Expected cost serves as the benchmark against which observed cost is compared to calculate the O/E ratio.
Hierarchical Condition Categories (HCCs)	A risk adjustment methodology that uses diagnosis codes to predict healthcare costs. HCCs group clinically related conditions and assign risk scores based on expected resource use.
Interquartile Range (IQR)	The difference between the 75th percentile and 25th percentile of a distribution. The IQR represents the range containing the middle 50% of values.
Merit-based Incentive Payment System (MIPS)	A program established by the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) that adjusts Medicare payments to clinicians based on performance in quality, cost, improvement activities, and Promoting Interoperability.
National Provider Identifier (NPI)	A unique 10-digit identification number assigned to healthcare providers. Used in combination with TIN to identify individual clinicians (TIN-NPI).
Observed Cost	The actual, payment standardized, Medicare payments made for services during an episode, before any risk adjustment.
Observed-to-Expected Cost Ratio (O/E Ratio)	The ratio of actual (observed) episode costs to the costs predicted (expected) by the risk adjustment model. An O/E ratio above 1.0 indicates higher-than-expected costs; below 1.0 indicates lower-than-expected costs.
Predictive Ratio	The ratio of average expected cost to average observed cost for a group of episodes. Used to assess calibration. Values close to 1.0 indicate good calibration.

<sup>13</sup> These documents will be available on the CMS Cost Measures Information page once field testing begins: <https://www.cms.gov/medicare/quality/value-based-programs/cost-measures/current>.

Term	Definition
Reliability	A measure of how consistently a measure differentiates performance between clinicians. Expressed as a value between 0 and 1, where higher values indicate that observed differences in scores more likely reflect true differences in performance rather than random variation.
Risk Adjustment	A statistical method that accounts for differences in patient characteristics (such as age, health status, and comorbidities) that may affect costs but are outside the clinician's control.
Risk-Adjusted Cost	Episode cost after applying the risk adjustment model to account for patient characteristics and other factors.
R-squared ( $R^2$ )	A statistical measure indicating the proportion of variance in episode costs explained by the risk adjustment model. Values range from 0 to 1, with higher values indicating the model explains more of the variation in costs.
Signal-to-Noise Ratio	A measure used to estimate reliability. "Signal" represents true differences in clinician performance; "noise" represents random variation within a clinician's episodes.
Stratification	The division of episodes into sub-groups based on patient or clinical characteristics to improve the accuracy of expected cost calculations.
Tax Identification Number (TIN)	A number used to identify a business entity (such as a group practice) for tax purposes. Used to identify clinician groups in measure reporting.
TIN-NPI	A combination of Tax Identification Number and National Provider Identifier used to identify an individual clinician practicing within a specific group.
Validity	The extent to which a measure accurately captures the concept it is intended to measure. For cost measures, validity assesses whether the measure reflects costs related to treatment choices and outcomes.

## Appendix B. Distributions of Measure Score

**Figure B1: Distribution of Measure Score - Group**



**Figure B2: Distribution of Measure Score - Clinician**

