

Non-Pressure Ulcers

Measure Testing Form

January 2026



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
	2.4.1 Conceptual Model.....	6
	2.4.2 Empirical Validity Analyses.....	7
2.5	Performance Gap.....	10
2.6	Risk Adjustment and Stratification	10
	2.6.1 Discrimination.....	11
	2.6.2 Calibration.....	11
2.7	Dual Status and Provider Location Analysis.....	12
2.8	Impact of Exclusions	16
	Appendix A. Glossary	18
	Appendix B. Distributions of Measure Score	20

1.0 Introduction

This Measure Testing Form (MTF) provides a summary of the preliminary measure testing results as part of field testing the Non-Pressure Ulcers 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).¹ 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.²

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.³

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

Non-Pressure Ulcers 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).

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

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

³ 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. Sections 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 Non-Pressure Ulcers 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	293,342
Mean Age	76
Part D Enrollment %	81.21%

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	3,730	100.00%	3,809	100.00%
Number of Episodes Attributed	-	-	-	-
20-39 Episodes	1,804	48.36%	2,704	70.99%
40-59 Episodes	717	19.22%	654	17.17%
60-79 Episodes	404	10.83%	247	6.49%
80-99 Episodes	249	6.67%	97	2.55%
100-199 Episodes	360	9.65%	98	2.57%
200-299 Episodes	93	2.49%	8	0.21%
300+ Episodes	103	2.76%	1	0.03%
Census Region	-	-	-	-
Northeast	721	19.33%	720	18.91%
Midwest	842	22.57%	796	20.89%
South	1,444	38.71%	1,599	41.97%
West	718	19.25%	691	18.14%
Unknown	5	0.13%	3	0.08%

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 non-pressure ulcers, including podiatrists, nurse practitioners, and family practitioners. These clinicians are also consistent with input provided by stakeholders, including patient and family partners (PFPs), during the measure development process. PFPs identified primary care providers, podiatrists, nurse practitioners, and surgeons as being part of their care team. Wound care specialists were also noted as an important specialty in the treatment and management of ulcers by the Clinical Expert Workgroup, but there is no specific CMS specialty designation for wound care at this time and these specialists will fall into existing designations.

Table 3: Count of the Top 10 Attributed Specialties

Specialty	Number of Clinicians Attributed
Podiatry	1,518
Nurse Practitioner	742
Family Practice	343
General Surgery	327
Internal Medicine	163
Physician Assistant	124
Emergency Medicine	121
Vascular Surgery	106
Plastic and Reconstructive Surgery	50
Infectious Disease	47

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	6,098	0.80	93.75%	8,951	0.76	90.48%
20	3,730	0.85	98.34%	3,809	0.83	97.30%
30	2625	0.89	99.73%	2,000	0.87	99.30%

At the testing volume of 20 episodes, the mean reliability for the Non-Pressure Ulcers measure is high, specifically 0.85 at the group level and 0.83 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.^{4,5} Nearly all 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 due to amputations, emergency room visits for cellulitis, 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).

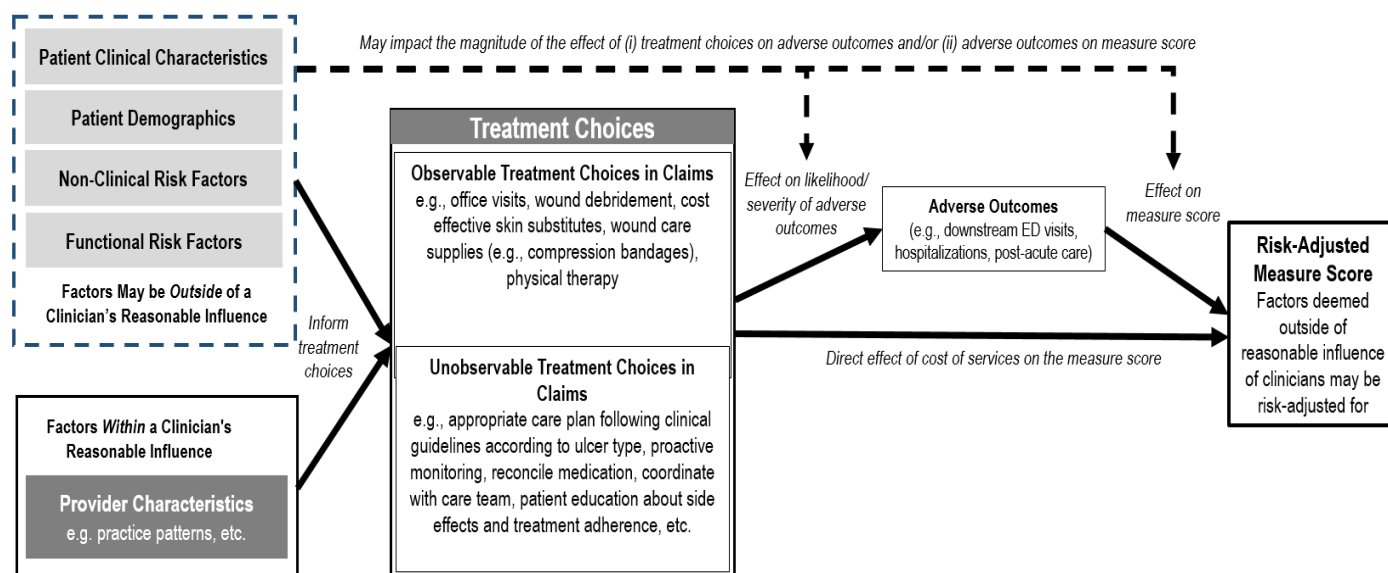
2.4.1 Conceptual Model

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.

⁴ 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.

⁵ 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



Patient-level attributes and preexisting risk factors falling both inside and outside a clinician's reasonable influence establish the clinical context in which clinicians determine the treatment choices along with their intensity and frequency delivered for the treatment and management of Non-Pressure Ulcers. These choices can be observable by the measure in claims data such as the use of related services like wound debridement and physical therapy as well as unobservable such as coordination with care teams. Yet, both types of choices affect the likelihood and severity of adverse outcomes. For instance, if a clinician provided wound debridement for a diabetic ulcer and subsequently coordinated care with a patient's physical therapist to provide offloading treatment and followed recommended clinical guidance, this would decrease the instance of adverse outcomes and their costs, such as an amputation due to a severe diabetic ulcer that was left untreated. Since there are factors falling outside a clinician's reasonable influence on adverse outcomes, such as complicating conditions like lymphedema and ulcer severity, the measure risk adjusts for these.

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%	\$5,213	\$39,153	\$3,783	\$7,434
Outpatient E/M Services	98.21%	\$5,258	\$39,399	\$3,812	\$7,452
Adverse Outcomes	-	-	-	-	-
Inpatient Hospital	5.44%	\$25,418	\$46,922	\$17,510	\$14,546
Major Procedures	4.38%	\$16,059	\$35,407	\$12,551	\$12,137
Emergency E/M Services	4.87%	\$17,008	\$54,407	\$10,761	\$12,440
Skilled Nursing Facility	2.17%	\$26,680	\$65,243	\$17,553	\$16,558

Table 5 demonstrates that Non-Pressure Ulcers episodes with adverse outcomes are substantially more costly than episodes overall. Episodes with inpatient hospitalizations have a mean observed cost of \$25,418, nearly five times higher than the mean observed cost of \$5,213 for all episodes. Similarly, episodes involving skilled nursing facility stays (\$26,680), emergency department visits (\$17,008), and major procedures such as amputations (\$16,059) all have mean observed costs that are three to five times higher than the baseline.

After risk adjustment, episodes with adverse outcomes continue to show higher costs than all episodes, though the difference is attenuated. For example, the mean risk-adjusted cost for episodes with inpatient hospitalizations is \$17,510, compared to \$3,783 for all episodes. This pattern, where adverse outcome episodes remain costlier after risk adjustment but the gap narrows, demonstrates that the measure appropriately accounts for patient-level factors that may predispose certain patients to adverse outcomes while still reflecting the cost impact of these events on the measure score.

These results support the validity of the Non-Pressure Ulcers cost measure by demonstrating that adverse outcomes, which represent potentially avoidable complications of ulcer management, meaningfully contribute to measure performance. Clinicians whose treatment choices reduce the likelihood of hospitalizations, emergency visits, skilled nursing facility stays, and major procedures such as amputations would be expected to have lower measure scores, consistent with the measure's intent to incentivize high-value care. 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 Services	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	0.13 [0.11,0.14] (p < 0.01)	-	0.12 [0.11,0.14] (p < 0.01)	-
Ambulatory/Minor Procedures	0.13 [0.10,0.16] (p < 0.01)	0.14 [0.08,0.20] (p < 0.01)	0.13 [0.10,0.15] (p < 0.01)	0.21 [0.15,0.27] (p < 0.01)
Durable Medical Equipment and Supplies	0.02 [0.02,0.02] (p < 0.01)	0.02 [0.01,0.02] (p < 0.01)	0.02 [0.02,0.02] (p < 0.01)	0.01 [0.00,0.01] (p < 0.01)
Laboratory, Pathology, and Other Tests	4.83 [4.00,5.67] (p < 0.01)	-0.88 [-2.57,0.82] (p = 0.31)	5.67 [4.98,6.37] (p < 0.01)	-0.26 [-1.75,1.24] (p = 0.73)
Major Procedures	0.20 [0.16,0.24] (p < 0.01)	0.17 [0.09,0.25] (p < 0.01)	0.23 [0.18,0.27] (p < 0.01)	0.11 [0.00,0.21] (p = 0.04)
Outpatient Evaluation & Management Services	-0.15 [-0.26,-0.03] (p = 0.01)	2.42 [2.20,2.64] (p < 0.01)	-0.10 [-0.23,0.03] (p = 0.13)	2.29 [2.02,2.55] (p < 0.01)
Part-D Drugs	0.27 [0.10,0.44] (p < 0.01)	0.68 [0.34,1.01] (p < 0.01)	0.36 [0.22,0.51] (p < 0.01)	0.61 [0.30,0.93] (p < 0.01)

The testing results in Table 6 demonstrate that the Non-Pressure Ulcers measure score reflects the cost directly related to treatment choices. Model 1 show that adverse events are statistically significantly associated with worse measure performance, confirming that episodes with adverse outcomes result in higher O/E ratios. Outpatient evaluation and management (E/M) services are associated with a statistically significant negative impact on the measure O/E ratios at the group level, though this relationship is not statistically significant at the clinician level. This suggests office visits may contribute to improved measure performance, potentially by facilitating earlier clinical intervention, and proactive management of patients' conditions. Laboratory, pathology, and other tests show the largest direct effect on measure performance at both group and clinician level, likely reflecting the extensive diagnostic workup required for clinically complex patients undergoing vascular procedures and related admissions. Other services including ambulatory/minor procedures, major procedures, and Part D drugs were only associated with a slightly worse performance at the clinician and group level, reflecting the direct cost contribution on care management of non-pressure ulcers.

Moreover, the indirect pathway, through which treatment choices influence the cost of adverse events, reveals important relationships that support measure validity. Model 2 shows that outpatient E/M services has the largest positive indirect impact on higher adverse event costs. This pattern suggests that patients with more frequent outpatient visits may have more complex ulcers requiring closer monitoring and treatment, and are at higher baseline risk for adverse outcomes rather than the visits themselves being the cause of adverse events. Alternatively, patients may require increased frequency of E/M services following an adverse event. Similarly,

Part-D Drugs also demonstrate significant positive indirect effects, with higher pharmaceutical utilization is associated with higher adverse event costs, suggesting that patients experiencing adverse events may have higher needs for Part D drugs.

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	\$4,363	\$3,998
Score Interquartile Range (IQR)	\$2,293	\$2,299
Standard Deviation	\$2,404	\$2,387
Coefficient of Variation	0.55	0.60
Score Percentile	Group	Clinician
10 th	\$2,129	\$1,835
25 th	\$2,909	\$2,502
50 th	\$3,854	\$3,479
75 th	\$5,201	\$4,801
90 th	\$7,066	\$6,692

Substantial variation is observed in the measure, indicated by the interquartile range, standard deviation, and coefficient of variation. The 90th percentile score is more than triple the 10th percentile score at the group level (\$2,129 vs \$7,066) and at the clinician level (\$1,835 vs \$6,692). The variation in the measure score, indicated by the interquartile range and standard deviation, is in the thousands of dollars. The results highlight an opportunity for improvement by closing the gap between the most and least efficient providers.

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.^{6,7,8,9} 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

⁶Assistant 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.

⁷Chen 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

⁸Medicare 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/>.

⁹Office 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>

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 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 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 0.185, and 0.181 after adjusting for the model's complexity based on the number of risk adjustors used. In other words, 18% 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. 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	1.16
Decile 2	1.06
Decile 3	1.00
Decile 4	0.92
Decile 5	0.93

Decile	Average Predictive Ratio
Decile 6	0.96
Decile 7	1.01
Decile 8	1.01
Decile 9	0.98
Decile 10	1.03

Table 8 shows that the measure's risk adjustment model shows moderate variation across risk deciles, with the range between 0.92 and 1.16, and an overall average of 1.00. For most risk deciles, the average predictive ratio is close to 1.00.

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.¹⁰ 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.

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"> Available for all beneficiaries Most powerful predictor of poor outcomes¹¹ 	<ul style="list-style-type: none"> Variation in Medicaid eligibility across states 	Yes
Provider Location (i.e., urban or rural)	<ul style="list-style-type: none"> Urban or rural provider location can reflect potential differences in patient access or barriers to care 	<ul style="list-style-type: none"> Reflects provider location, not beneficiary residence, which may differ 	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

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

¹¹ Refer to footnote 4.

locality 'rural' designation using the most recent quarter of mapping data available for the 2024 study period or any year within a 2 year lookback window. If the average urban and rural providers' scores are close to each other and to the national average provider score, it 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 Scores (Percentiles)				
				P10	P25	P50	P75	P90
Group	All Providers	100.00%	\$4,363	\$2,129	\$2,909	\$3,854	\$5,201	\$7,066
	Urban	83.73%	\$4,398	\$2,130	\$2,932	\$3,879	\$5,236	\$7,177
	Rural	16.27%	\$4,179	\$2,123	\$2,802	\$3,746	\$4,999	\$6,492
Clinician	All Providers	100.00%	\$3,998	\$1,835	\$2,502	\$3,479	\$4,801	\$6,692
	Urban	85.67%	\$4,028	\$1,801	\$2,508	\$3,491	\$4,846	\$6,733
	Rural	14.33%	\$3,821	\$1,949	\$2,464	\$3,365	\$4,614	\$6,128

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.

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 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	Arterial Ulcer Type without Part D	1.03%	0.01 (p = 0.97)	-0.02 (p = 0.93)	0.20 (p = 0.39)
	Arterial Ulcer Type with Part D	4.43%	-0.04 (p = 0.27)	-0.07 (p = 0.05)	-0.10 (p = 0.02)
	Diabetic Ulcer Type without Part D	6.02%	0.24 (p = 0.00)	0.23 (p = 0.00)	0.25 (p = 0.00)
	Diabetic Ulcer Type with Part D	27.52%	0.07 (p < 0.0001)	0.05 (p = 0.00)	0.04 (p = 0.01)
	Venous Ulcer Type without Part D	4.10%	0.21 (p = 0.04)	0.17 (p = 0.09)	0.14 (p = 0.25)
	Venous Ulcer Type with Part D	16.13%	0.11 (p < 0.0001)	0.08 (p < 0.0001)	0.10 (p < 0.0001)
	Multiple Ulcer Types without Part D	2.69%	0.24 (p = 0.04)	0.26 (p = 0.02)	0.10 (p = 0.49)
	Multiple Ulcer Types with Part D	11.54%	0.08 (p = 0.00)	0.04 (p = 0.05)	0.02 (p = 0.30)
	Non-Specific Ulcer Type without Part D	5.00%	0.01 (p = 0.84)	0.00 (p = 0.97)	-0.07 (p = 0.41)
	Non-Specific Ulcer Type with Part D	21.53%	-0.02 (p = 0.25)	-0.02 (p = 0.13)	-0.02 (p = 0.26)
Clinician	Arterial Ulcer Type without Part D	1.04%	0.00 (p = 1.00)	-0.03 (p = 0.85)	0.57 (p = 0.12)
	Arterial Ulcer Type with Part D	4.46%	-0.04 (p = 0.22)	-0.08 (p = 0.06)	-0.10 (p = 0.07)
	Diabetic Ulcer Type without Part D	6.01%	0.24 (p = 0.00)	0.26 (p = 0.00)	0.39 (p = 0.00)
	Diabetic Ulcer Type with Part D	27.47%	0.08 (p < 0.0001)	0.05 (p = 0.00)	0.05 (p = 0.01)
	Venous Ulcer Type without Part D	4.09%	0.26 (p = 0.01)	0.23 (p = 0.03)	0.32 (p = 0.06)
	Venous Ulcer Type with Part D	16.10%	0.11 (p < 0.0001)	0.08 (p = 0.00)	0.10 (p < 0.0001)
	Multiple Ulcer Types without Part D	2.68%	0.20 (p = 0.09)	0.21 (p = 0.08)	0.15 (p = 0.50)
	Multiple Ulcer Types with Part D	11.51%	0.09 (p < 0.0001)	0.07 (p = 0.00)	0.04 (p = 0.14)
	Non-Specific Ulcer Type without Part D	5.00%	0.05 (p = 0.47)	0.07 (p = 0.33)	0.04 (p = 0.68)
	Non-Specific Ulcer Type with Part D	21.65%	-0.02 (p = 0.24)	-0.02 (p = 0.15)	-0.01 (p = 0.45)

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	1.15	1.15	1.14	1.08	1.07	1.05
0-20%	1.10	1.03	1.09	1.02	0.90	1.02
21-40%	1.15	1.19	1.14	1.05	1.03	1.05
41-60%	1.12	1.15	1.11	1.05	1.05	1.04
61-80%	1.14	1.12	1.14	1.09	1.11	1.06
81-100%	1.26	1.25	1.21	1.17	1.21	1.08

Table 13: Proportions of Groups and Clinicians Who Perform Significantly Worse, Equally Well, or Significantly Better on Their Dual Episodes than Non-Dual Episodes

Provider Level	Significantly Worse	Equally Well	Significantly Better
Group	7.30%	90.37%	2.33%
Clinician	8.26%	88.90%	2.84%

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

Provider 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	38.07%	0.67%
Clinician	39.93%	0.60%

These results suggest that risk adjustment for patient-level dual status remains appropriate for this measure, while clinician-level dual share does not substantially improve the model. Table 11 shows that the association between patient-level dual status and episode cost remains statistically significant for several major subgroups even after controlling for clinician-level factors. Most notably, Diabetic Ulcer Type subgroups—representing over 33% of all episodes—demonstrate a consistent and significant association across all model specifications. The Venous Ulcer Type with Part D subgroup (approximately 16% of episodes) shows a similar pattern. Importantly, adding clinician's dual share or clinician fixed effects does not eliminate these patient-level associations. While coefficients attenuate slightly in some subgroups, the fundamental relationship persists. This indicates that cost differences associated with dual-eligible patients are primarily driven by patient-level characteristics rather than clinician practice patterns or provider's patient composition. The persistence of significant patient-level effects in the largest subgroups supports continuing to risk adjust dual at the patient level.

Additionally, Table 12 demonstrates that measure performance degrades and becomes more variable as clinician and group dual share increases, for both dual and non-dual episodes. Across many deciles, performance is worse for dual episodes, especially at the clinician level. Approximately 90% of groups and clinicians perform equally well on both dual and non-dual episodes. However, more clinicians and groups perform worse on their dual episodes than

perform better, suggesting that providers aren't able to fully mitigate the effect of SRFs (Table 13). Lastly, risk adjusting for dual status appears to change measure performance for a subset of clinicians, but only less than 1 percent of groups and clinician would see a ranking shift by 5% or more (Table 14).

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 th	25 th	50 th	75 th	90 th
All Episodes Meeting Triggering Logic	430,597	100.00%	\$5,359	\$179	\$316	\$901	\$4,055	\$11,786
Episode Length Less Than 90 Days	26,044	6.05%	\$11,978	\$466	\$847	\$1,722	\$4,550	\$14,784
Beneficiary Death in Episode	47,846	11.11%	\$12,557	\$434	\$852	\$1,997	\$6,060	\$18,292
Outlier Cases	7,218	1.68%	\$22,765	\$317	\$1,169	\$24,725	\$45,937	\$46,428
No Attributed Clinician (Clinician Reporting Only)	16,755	3.89%	\$7,832	\$445	\$906	\$2,572	\$7,633	\$18,693
Recent Hospice Use	7,717	1.79%	\$12,376	\$150	\$290	\$808	\$2,110	\$6,114
Group does not Meet Testing Volume Threshold	90,167	20.94%	\$6,798	\$177	\$282	\$764	\$4,708	\$14,353
Clinician does not Meet Testing Volume Threshold	236,544	54.93%	\$5,879	\$179	\$301	\$832	\$4,411	\$13,116
Calcinosis Cutis	654	0.15%	\$13,490	\$187	\$510	\$2,482	\$8,445	\$20,804
Calciophylaxis	1,754	0.41%	\$14,212	\$272	\$748	\$2,557	\$9,643	\$28,839
Ulcers Associated With Fistulae	10,836	2.52%	\$8,904	\$181	\$378	\$1,247	\$5,541	\$15,488
Hidradenitis Suppurativa	660	0.15%	\$4,719	\$174	\$340	\$852	\$2,869	\$8,838
Pyoderma Gangrenosum	1,706	0.40%	\$20,640	\$218	\$672	\$3,090	\$8,557	\$20,621
Scleroderma	1,351	0.31%	\$7,085	\$196	\$357	\$1,131	\$5,630	\$14,513
Sickle Cell Anemia	456	0.11%	\$17,477	\$204	\$405	\$1,475	\$8,011	\$19,773
Vasculitis	4,369	1.01%	\$10,029	\$211	\$445	\$1,713	\$6,506	\$15,907

Exclusion Criteria	Episodes		Observed Episode Cost					
	Count	Percent of All Episodes Meeting Trigger Logic	Mean	Percentile				
				10 th	25 th	50 th	75 th	90 th
Reportable Episodes (if all clinicians reported as Group at the Testing Volume Threshold)	277,776	64.51%	\$3,686	\$172	\$299	\$774	\$3,423	\$9,710
Reportable Episodes (if all clinicians reported as Clinician at the Testing Volume Threshold)	144,378	33.53%	\$3,281	\$163	\$289	\$738	\$2,876	\$8,480

Table 15 displays descriptive statistics of all episodes meeting the measure's triggering logic, excluded episodes, and the final reportable episodes at the group-and individual level. The statistical results show that the reportable episode populations are more homogenous and comparable than all episodes meeting triggering logic since excluding these episodes reduces the variation in the distribution of observed costs. For instance, the observed cost of all episodes meeting trigger logic had a mean observed cost of \$5,359 and ranged from \$179 in the 10th percentile to \$11,786 in the 90th percentile. Meanwhile, the reportable episodes at the group level had a mean observed cost of \$3,686 with a smaller distribution ranging from \$172 in the 10th percentile to \$9,710 in the 90th percentile. Similarly, the reportable episodes at the TIN-NPI had a mean observed cost of \$3,281 ranging from \$163 in the 10th percentile to \$8,480 at the 90th percentile. With the exception of the Hidradenitis Suppurativa excluded cohort, which only comprises 0.15% of triggered episodes, the excluded cohorts have higher mean observed costs compared to all episodes meeting triggering logic. The largest exclusions come from applying the case minimum threshold of 20 episodes. Overall, these findings support the exclusion of these episodes to ensure a comparable patient cohort that will yield a clinically coherent measure and meaningful information to attributed 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. ¹²
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.

¹² 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

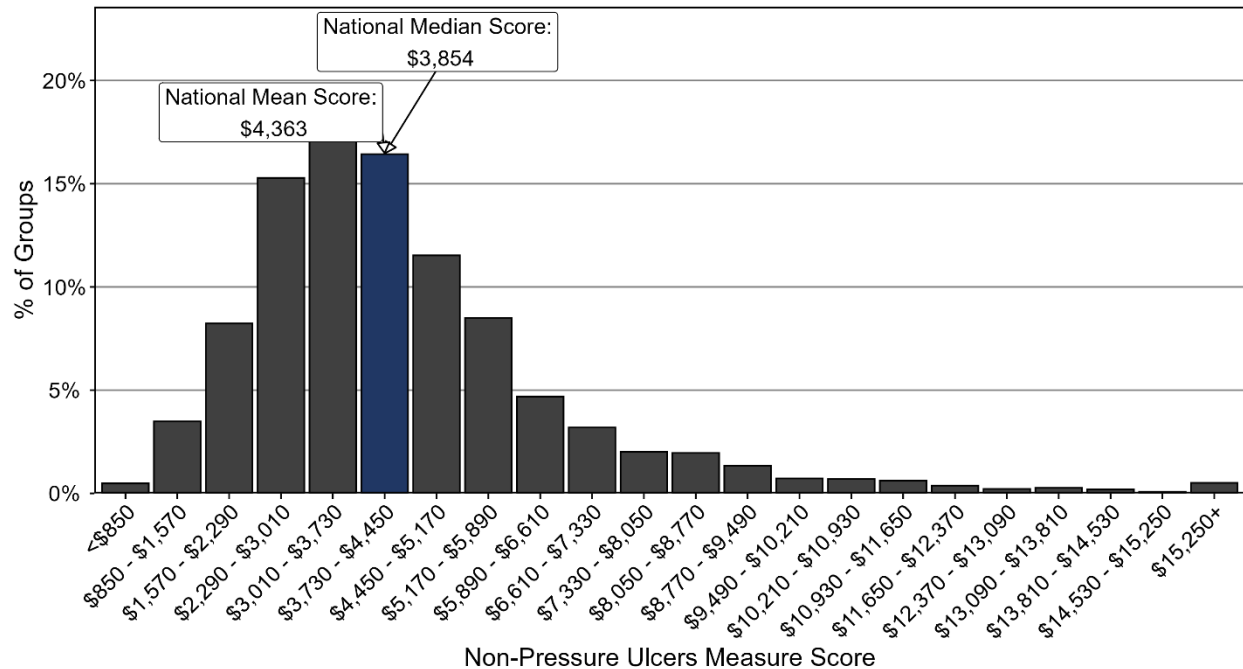


Figure B2: Distribution of Measure Score - Clinician

