Quality Payment
PROGRAM

Emergency Medicine

Measure Testing Form

January 2022

Contents

1.0 Introduction

This Measure Testing Form (MTF) provides a brief summary of the preliminary measure testing results as part of field testing five episode-based cost measures. Stakeholders may review these results, alongside other documentation, to provide feedback on the draft measure using the field testing survey. The testing results reflect the performance of the measure as specified at the time of field testing, which is part of the measure development process. Please see the Draft Cost Measure Methodology for a description of the measure specifications and the Draft Measure Codes List for the list of codes used to specify the measure.[1]

1.1 Project Title and Overview

The Centers for Medicare & Medicaid Services (CMS) has contracted with Acumen, LLC to develop care episode and patient condition groups for use in cost measures to meet the requirements of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). The contract name is "Physician Cost Measures and Patient Relationship Codes (PCMP)." The contract number is 75FCMC18D0015, Task Order 75FCMC19F0004.

1.2 Measure Name

Emergency Medicine

1.3 Type of Measure

Cost/Resource Use

1.4 Data

The study period is January 1, 2019 through December 31, 2019. All episodes ending during the study period that meet inclusion and exclusion criteria are included in testing. The measure is calculated with Medicare Parts A, B, and D administrative claims data, Long-Term Minimum Data Set, Medicare Enrollment Database. For testing purpose, other data sources are used, including the American Community Survey, Common Medicare Environment, and Uniform Data System.

Testing results are presented at a testing volume threshold of 20 episodes for clinician groups and individual practitioners. Clinician groups are identified by a Tax Identification Number (TIN). Individual clinician are identified using a combination of a Tax Identification Number and National Provider Identifier (TIN-NPI).

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

2.0 Preliminary Testing Results

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

2.1 Measure Coverage

Table 1 shows the number of patients and percent of Medicare Parts A and B costs covered by this measure. This measure has the potential to have high impact as a large number of patients receive emergency medicine care, representing a substantial share of Medicare spending.

Table 1. Beneficiary and Cost Coverage

Coverage Metrics	Testing Volume Threshold	
	1 Episode	20 Episodes
Number of Patients	8,168,166	8,165,332
Percent of Parts A & B Costs – TIN	23.0%	23.0%
Percent of Parts A & B Costs – TIN-NPI	23.0%	22.8%

2.2 Frequently Attributed Specialties

Table 2 shows the top 10 attributed specialties for this measure, using a 20-episode testing volume threshold. As intended, the measure primarily includes emergency medicine clinicians. Other specialties listed in Table 2 include a range of clinicians that provide care, such as specialized diagnostic or treatment services, in the emergency department. These specialties are also consistent with input provided by stakeholders, including Person and Family Partners (PFPs), during the measure development process. PFPs identified emergency medicine clinicians, nurses, specialists such as cardiologists, and surgeons amongst others as being part of their care team during the emergency department visit.

Table 2 also provides metrics to show the breakdown of the share of episodes covered by each specialty. Emergency medicine clinicians make up the majority of all clinicians who meet the testing volume threshold (65.8%). Physician assistants and nurse practitioners are the second and third most frequently attributed specialties, comprising 16.1% and 9.2% of all attributed clinicians, respectively.

Finally, Table 2 shows the percentage of each specialty that is covered by this measure. 68.0% of all emergency medicine clinicians who billed at least one Part B physician/supplier claim in 2019 have at least 20 Emergency Medicine episodes, reflecting the measure's focus on this type of care. Note that this metric reflects the variation in size across specialties; for instance, the share of nurse practitioners appears low (3.7%) but it is a very large specialty.

For more information about the measure's attribution methodology, please refer to the Draft Cost Measure Methodology.

Table 2. Count and Attribution Frequency of the Top 10 Attributed Specialties at Testing Volume Threshold of 20 Episodes

Rank	Specialty	Number of TIN-NPIs Attributed	Percent of Specialty Among All Attributed TIN-NPIs	Percent of All Episodes	Percent of Specialty (TIN-NPI) Attributed to Measure
1	Emergency Medicine	52,314	65.8%	77.9%	68.0%
2	Physician Assistant	12,798	16.1%	10.0%	11.0%
3	Nurse Practitioner	7,288	9.2%	5.6%	3.7%
4	Family Practice	4,264	5.4%	4.5%	3.7%
5	Internal Medicine	1,798	2.3%	2.0%	1.3%
6	General Surgery	420	0.5%	0.2%	1.5%
7	General Practice	414	0.5%	0.3%	5.2%
8	Psychiatry	322	0.4%	0.1%	0.9%
9	Cardiology	197	0.3%	0.1%	0.6%
10	Neurology	194	0.2%	0.1%	1.0%

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 3 shows reliability metrics at various testing volume thresholds. While higher thresholds yield higher reliability results, it is at the cost of further reducing the number of clinicians and clinician groups eligible for the measure, which would reduce the potential impact of the measure. For the purposes of field testing, we used a 20-episode volume threshold (bolded in the table below); for simplicity, we use this threshold across all measures. 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 3. Sample Size, Mean Reliability, and Proportion of Clinicians above Moderate Reliability at Various Testing Volume Thresholds

Testing Volume Threshold	TIN			TIN-NPI		
	Number of TINs	Mean Reliability	Percent Above 0.4	Number TIN-NPIs	Mean Reliability	Percent Above 0.4
10	5,111	0.81	96.5%	90,924	0.73	89.4%
20	**4,071**	**0.90**	**100.0%**	**79,540**	**0.79**	**100.0%**
30	3,617	0.93	100.0%	72,287	0.82	100.0%

At the testing volume of 20 episodes, the measure achieved high reliability at both the TIN and TIN-NPI levels, at 0.90 and 0.79 respectively (Table 3). CMS generally considers 0.4 as the threshold indicating 'moderate' reliability and 0.7 as 'high' reliability, which is supported by previous work into reliability and the threshold was finalized in the CY 2022 Physician Fee

Schedule final rule.[2,3] Additionally, all providers meet or exceed the moderate reliability threshold of 0.4.

2.4 Validity

Validity is a criterion that evaluates whether the cost measure is able to quantify the construct that it aims to measure, which is the cost performance of clinicians. Validity is tested empirically by examining the association between the measure score and high-cost events that drive the measure score, such as downstream complications and consequences of care, and the correlation between cost and quality measures.

The measure score reflects the average risk-adjusted cost of episodes that were attributed to a particular clinician or group. The risk-adjusted cost neutralizes the effects of risk factors deemed to be outside of a clinician's influence (e.g., pre-existing conditions, age, or indicators of clinical severity) from the standardized cost[4] of clinically relevant services observed during a care episode. For the full list of factors that were risk adjusted for this measure, please see the Draft Cost Measure Methodology.

Table 4 shows that episodes that had another ED visit (or return visit) after the initial ED visit had higher mean observed and risk-adjusted costs than the overall population of episodes included in the measure. This observed association is consistent with a return visit being a downstream high-cost event that can drive up cost if not avoided. The results show that the cost measure is able to differentiate the cost efficiency of episodes based on high-cost events.

Table 4: Distribution Episode Cost Stratified by High-Cost Events

High-Cost Events	Observed Cost			Risk-Adjusted Cost		
	Mean	Standard Deviation	Median	Mean	Standard Deviation	Median
All Episodes	$7,242	$9,016	$3,286	$7,029	$7,444	$5,253
Emergency Department Return Visit	$8,146	$8,526	$4,817	$9,944	$9,148	$7,294

We also examined the correlation between the cost measure and related quality measures to test the relationship between these different metrics. While there are important limitations to this analysis, it can provide some indication of whether there is variation in cost with different levels of quality. In brief, we note the following key points for interpreting the strength and direction of correlations:

---

[2] Mathematica, Inc., "Memorandum: Reporting Period and Reliability of AHRQ, CMS 30-Day and HAC Quality Measures – Revised," http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability-.pdf.

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

[4] Claim payments are standardized to account for differences in Medicare payments for the same service(s) across Medicare providers. Payment standardized costs remove the effect of differences in Medicare payment among health-care providers that are the result of differences in regional health-care provider expenses measured by hospital wage indexes and geographic price cost indexes (GPCIs) or other payment adjustments, such as those for teaching hospitals. For more information, please refer to the "CMS Price (Payment) Standardization - Basics" and "CMS Price (Payment) Standardization - Detailed Methods" documents posted on the Payment Standardization webpage. (https://resdac.org/articles/cms-price-payment-standardization-overview)

- A strong inverse correlation – good performance on cost with poor quality performance – would indicate that variation in cost is solely reflective of variation in quality. This suggests that care stinting could be a concern.
- A weak correlation between cost and quality in either a positive or negative direction indicates variation in cost at any given level of quality. This suggests that cost performance can be improved without negatively impacting quality.

In general, the direction of correlations indicate the following:

- Positive correlations with quality measures indicate that clinicians providing better quality care on that particular metric tend to also have lower costs. That is, clinicians who have high rates of performing specific quality actions (as measured through process measures) or achieve better patient health outcomes (as measured through outcomes measures) tend to have lower costs of care. As such, these associations could represent ways to lower costs while also providing high-quality care
- A negative correlation between a cost and quality measure does not indicate an absence of cost improvement potential consistent with high-value care. This is because other approaches to improving cost performance (e.g., patient education) may not be captured by the selected quality measure.

There are several key considerations regarding the conceptual relationship between measures and data limitations when interpreting the results. The extent to which correlations can provide meaningful information depends on what is being measured. Ideally, measures should apply to the same care provided for the same patient cohort for the same time horizon. Correlations with a quality measure that focuses on outcomes for the same patient cohort may be more informative than a broadly applicable process measure that applies across a wide range of conditions. Clinicians select only 6 MIPS quality measures to report and are generally required to only report one outcome or high-priority measure. This selective reporting likely biases the observed sample.

Table 5 shows the correlation between the Emergency Medicine cost measure and related MIPS quality measures. To aid interpretability, the direction of all non-inverse quality measures is inverted so that a lower score indicates better performance for all cost and quality measures. Quality measures are selected based on their clinical proximity to the cost measure, such as assessing quality actions related to a similar patient cohort, and the number of clinicians with both cost and quality measures.

There are very low numbers of TINs and TIN-NPIs that have both the quality and cost measure, providing a limited sample for this analysis. There are 4,071 TINs and 79,540 TIN-NPIs that meet the testing volume threshold for Emergency Medicine; across all the quality measure pairs tested, the highest number of overlapping providers was 14,452 TIN-NPIs. In general, the correlations are weakly positive, suggesting that cost performance can be improved without negatively impacting quality. One measure assessing time spent in the ED shows a statistically significant small to medium positive correlation with the cost measure at both the TIN and TIN-NPI levels (QECEP50: ED Median Time from ED Arrival to ED Departure for Discharged ED Patients for Adult Patients). The results suggest that providers who tend to perform well on the cost measure also tend to perform well on the quality metric. There are no strong inverse correlations, suggesting that care stinting is not a concern within the limitations of this analysis.

Table 5. Correlation between Quality Measures and Cost Measure

Related Quality Measure	TIN or TIN-NPI	Number of Entities with Both Cost and Quality Measures Available	Pearson Correlation	P-value
Q415: Emergency Medicine: Emergency Department Utilization of CT for Minor Blunt Head Trauma for Patients Aged 18 Years and Older (Efficiency)	TIN	447	0.02	0.61
	TIN-NPI	14,452	0.01	0.12
Q458: All-cause Hospital Readmission (Outcome)^	TIN	344	-0.08	0.12
QACEP50: ED Median Time from ED Arrival to ED Departure for Discharged ED Patients for Adult Patients (Process)	TIN	98	0.26	0.01*
	TIN-NPI	4,038	0.17	0.00*

*P-value <0.05 indicates statistical significance
^Replaced from MIPS CY 2021 onwards with a re-specified version: Q479 Hospital-Wide, 30-Day, All-Cause Unplanned Readmission (HWR) Rate for MIPS Groups

2.5 Performance Gap

Table 6 shows the distribution of the measure score for TIN and TIN-NPI levels. The score interquartile ranges, standard deviations, and coefficients of variation are similar for the TIN and TIN-NPI distributions. The 90th percentile is about 1.5 times higher than the 10th percentile of measure score for both TIN and TIN-NPI levels. The variation in the measure score are in the thousands of dollars, therefore there are meaningful opportunities to improve cost efficiency.

Table 6. Distribution of the Measure Score

Metric	TIN	TIN-NPI
Mean Score	$7,460	$7,105
Score Interquartile Range (IQR)	$1,198	$1,333
Standard Deviation	$1,503	$1,201
Coefficient of Variation	0.20	0.17
Score Percentile		
10th	$6,060	$5,616
25th	$6,720	$6,449
50th	$7,251	$7,163
75th	$7,917	$7,782
90th	$9,194	$8,413

2.6 Risk Adjustment and Stratification

2.6.1 Discrimination

Discrimination is a statistical criterion that evaluates the measure's ability to distinguish high-cost episodes from low-cost episodes, or the ability to explain the variance in cost of individual episodes. The amount of variance explained is estimated by the R-squared metric with the range between 0 and 1. The R-square for the measure is 0.53, and 0.53 after adjusting for the model's complexity based on the number of risk adjustors used. In other words, 53% of the

variation in the actual observed cost of episodes is explained by the risk adjustment model and sub-group stratification.

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

2.6.2 Calibration

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

Table 7 shows an overall consistency of predicted measure scores across levels of risk, with a range of average predicted ratios no more than 0.03 above and 0.02 below 1.00. These results suggest that the Emergency Medicine measure is well-calibrated across the full range of resource use patterns observed in the population.

Table 7. Predictive Ratio by Decile of Predicted Episode Cost

Decile	Average Predictive Ratio
Decile 1	1.01
Decile 2	1.02
Decile 3	1.03
Decile 4	1.03
Decile 5	1.01
Decile 6	0.99
Decile 7	0.98
Decile 8	0.99
Decile 9	0.99
Decile 10	1.00

2.7 Social Risk Factor Analysis

Beyond clinical characteristics of patients, the cost of care may be influenced by non-clinical factors related to a patient's social risk factors (SRFs), such as race, income, education, and employment. At the program level, MIPS adjusts for SRFs using the MIPS Complex Patient Bonus to ensure clinicians or groups treating more complex patients are not disadvantaged.[5]

[5] https://qpp-cm-prod-content.s3.amazonaws.com/uploads/966/QPP%20COVID-19%20Response%20Fact%20Sheet.pdf

To examine the extent that SRFs can mask the underlying performance of clinicians, this analysis added SRFs to the base risk adjustment model to examine their potential impact to clinicians' scores and goodness of fit of the risk adjustment model. The base risk adjustment model includes the standard set of risk adjustors from the CMS-Hierarchical Condition Categories (HCC) version 22 in 2016, disability status, End-Stage Renal Disease (ESRD) status, comorbidity interaction variables, recent long-term care use, HCC count, and measure-specific clinical risk adjustors. For the full list of factors that were risk adjusted for this measure, please see the Draft Cost Measure Methodology.

The base model was compared against 3 models that included additional SRFs from the American Community Survey, Common Medicare Environment described below:

1. Dual Medicare-Medicaid eligibility status:
    - Base model
    - Medicare-Medicaid eligibility status: full, partial, or non-dual status
2. All Socioeconomic Status (SES) Variables Model:
    - Base model
    - Dual Medicare-Medicaid eligibility status: full, partial, or non-dual status
    - Sex: female or male
    - Race: Asian, Black, Hispanic, North American Native, White, and other
    - Neighborhood Income Level: low income (median income < 33rd percentile nationally), medium income (median income between 33rd and 66th percentiles), and high income (median income > 66th percentile)
    - Neighborhood Education Rate: majority less than high school, high school, or greater than high school
    - Neighborhood Employment Rate: less than or greater than 10%
3. Agency for Healthcare Research and Quality SES Index Model:
    - Base model
    - Dual Medicare-Medicaid eligibility status
    - Sex: female or male
    - Race: Asian, Black, Hispanic, North American Native, White, and other
    - AHRQ SES Index Score: calculated using the AHRQ's scoring algorithm, based on data from the American Community Survey

Table 8 shows the percent of providers who would see their performance shifted, based on their percentile ranking, using the new SRF models. There are marginal changes to the performance at both TIN and TIN-NPI reporting levels compared to the base model. Across the 3 models, over 95% TINs and TIN-NPIs would experience small and likely inconsequential swing in their performance, of less than 5% compared to the base model. Additionally, the performance ranking has Spearman correlation values of greater than 0.99 between the base model and each of the SRF models, which also suggests that adjusting for SRF variables does not fundamentally change how the measure evaluates provider performance.

Table 8. Clinicians' Performance Shift Measured by the Change in the Average Ratio of Observed-to-Expected Cost

Model	TIN or TIN-NPI	Proportion of Clinicians Affected at Various Levels of Performance Shift						
		-10% or less	-9% to -6%	-5% to -2%	-1% to 1%	2% to 5%	6% to 9%	10% or more
Model 1	TIN	*	0.4%	8.3%	84.5%	6.7%	*	*
	TIN-NPI	0.1%	0.4%	9.2%	82.7%	7.4%	0.1%	0.0%
Model 2	TIN	*	0.7%	16.3%	67.8%	13.2%	1.4%	0.5%
	TIN-NPI	0.2%	1.1%	17.9%	64.2%	13.8%	1.9%	0.8%
Model 3	TIN	*	0.8%	17.0%	66.2%	13.7%	1.7%	0.5%
	TIN-NPI	0.2%	1.1%	18.1%	64.2%	13.6%	2.0%	0.9%

Note: Cells with stars ('*') indicate values that were suppressed as there were less than 10 observations.

2.8 Impact of Exclusions

Table 9 displays descriptive statistics of all episodes meeting the measure's triggering logic, excluded episodes, and final reportable episodes at both TIN and TIN-NPI levels. These exclusion criteria ensure that the reportable episode populations are more homogenous and comparable than all episodes meeting triggering logic. It is worth noting that only the observed cost is shown, which has not been risk adjusted for using our risk adjustment model. Therefore, the differences in cost may appear much smaller after risk adjustment than as-is.

Overall, the exclusion criteria decrease the mean episode cost by approximately $1,600 compared to the mean cost of all episodes meeting triggering logic, at $8,847 (Table 9).

Episodes with beneficiary death during an episode also have higher mean observed cost than all episodes meeting triggering logic, at $16,746 (Table 9). Excluding these episodes ensures that episodes are comparable and the observation window is not truncated.

Episodes that are not reliably predicted by the risk adjustment model are excluded because they deviate substantially from the projected cost for a given patient risk profile than the all episodes meeting triggering logic, with mean observed cost of $34,187 (Table 9).

Based on the input of stakeholders during the development process, the following episodes are excluded because they have different care pathways or characteristics that make them incomparable to the rest of the episodes: episodes with medical complications or cannot be classified into a visit type defined by the clinical expert workgroup, emergency department to emergency department transfers, and hospital to hospital transfers. More information on the definitions of these criteria can be found in the Draft Cost Measure Methodology. Except for episodes with hospital to hospital transfers, the mean observed costs for episodes with medical complications or cannot be classified into a visit type and ED to ED transfers are lower than that of all episode meeting triggering logic (Table 9).

The last exclusion criteria come from applying the testing volume threshold to ensure that there is sufficient sample size to calculate the measure for providers.

Table 9: Cost Statistics for Measure Exclusions

Exclusion Criteria	Episodes		Observed Episode Cost					
				Percentile				
	Count	Percent of All Episodes Meeting Trigger Logic	Mean	10th	25th	50th	75th	90th
All Episodes Meeting Triggering Logic	17,669,864	100.0%	$8,847	$672	$1,247	$3,788	$11,914	$23,458
Beneficiary Death in Episode	744,084	4.2%	$16,746	$2,726	$8,130	$12,913	$19,643	$35,008
Outlier Cases	311,006	1.8%	$34,187	$5,477	$11,233	$30,831	$46,755	$67,520
Episodes with Medical Complications or Cannot Be Classified into a Visit Type	873,909	5.0%	$7,380	$535	$1,000	$2,341	$8,438	$21,319
Emergency Department to Emergency Department Transfer	134,676	0.8%	$7,883	$1,557	$2,622	$4,830	$9,497	$17,736
Hospital to Hospital Transfer	436,499	2.5%	$39,542	$12,628	$21,892	$34,423	$47,738	$67,921
TIN does not Meet Testing Volume Threshold	7,485	0.0%	$9,628	$454	$1,101	$4,085	$11,745	$24,124
TIN-NPI does not Meet Testing Volume Threshold	237,570	1.3%	$6,848	$435	$831	$2,017	$8,195	$19,137
Reportable Episodes (if all clinicians reported as TIN at the testing volume threshold)	**15,234,645**	**86.2%**	**$7,242**	**$644**	**$1,171**	**$3,286**	**$10,086**	**$20,171**
Reportable Episodes (if all clinicians reported as TIN-NPI at the testing volume threshold)	**15,034,854**	**85.1%**	**$7,267**	**$649**	**$1,178**	**$3,309**	**$10,128**	**$20,220**