Quality Payment
PROGRAM

End-Stage Renal Disease

Measure Justification Form

December 2023

Table of Contents

1.0 Introduction

This Measure Justification Form (MJF) provides results for the testing and evaluation of the End-Stage Renal Disease (ESRD) measure. The form is intended to provide detailed information about the testing conducted on this measure, and accompanies the Measure Methodology[1] and Measure Codes List[2] file, which together, comprise the specifications for this cost measure.

1.1 Project Title

Physician Cost Measure and Patient Relationship Codes

1.2 Date

Information included is current on December, 8 2023

1.3 Project 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 Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) requirements. The contract name is "Physician Cost Measure and Patient Relationship Codes (PCMP)." The contract number is 75FCMC18D0015, Task Order 75FCMC19F0004.

1.4 Measure Name

End-Stage Renal Disease Episode-Based Cost Measure

1.5 Type of Measure

Cost/Resource Use

1.6 Measure Description

The ESRD episode-based cost measure evaluates a clinician's or clinician group's risk-adjusted and specialty-adjusted cost to Medicare for patients who receive medical care to manage ESRD. This chronic condition measure includes the costs of services that are clinically related to the attributed clinician's role in managing care during an ESRD episode.

[1]CMS, "End-Stage Renal Disease Measure Methodology," *QPP Cost Measure Information Page,* https://www.cms.gov/medicare/quality/value-based-programs/cost-measures
[2]CMS, "End-Stage Renal Disease Measure Codes List" *QPP Cost Measure Information Page*, https://www.cms.gov/medicare/quality/value-based-programs/cost-measures

2.0 Importance

2.1 Evidence to Support the Measure Focus

The ESRD measure was developed for use in the Merit-based Incentive Payment System (MIPS) to meet the requirements of the Social Security Act section 1848(r), added by the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). MIPS aims to reward high-value care by measuring clinician performance through four areas: quality, improvement activities, promoting interoperability, and cost. Each category assesses different aspects of care, and the categories are weighted to combine into one composite score. CMS introduced MIPS Value Pathways (MVPs) to align and connect quality measures, cost measures, and improvement activities across performance categories of MIPS for different specialties or conditions. MVPs aim to provide a holistic assessment of clinician value for a specific type of care to achieve better healthcare outcomes and lower patient costs.

The use of cost measures is required by statute, and their purpose is to assess resource use. To be effective, they should capture costs related to a clinician's care decisions and account for factors outside their influence. This measure provides clinicians with information about their care costs that they can use to understand the costs associated with their decision-making. Clinicians play an important role in variation in health care expenditures due to their ability to affect costs.[3] A cost measure offers an opportunity for improvement if clinicians can exercise influence on the intensity or frequency of a significant share of costs during the episode, or if clinicians can achieve lower spending and better quality of care quality through changes in clinical practice.

According to the literature and feedback received through stakeholder input activities, this measure's focus represents an area with opportunities for improvement. As discussed in the rest of this section, primary opportunities for improving ESRD cost outcomes include incentivizing appropriate dialysis care and improving care coordination.

Dialysis costs are the most significant cost for ESRD patients, but more importantly the form of dialysis care (i.e., peritoneal dialysis or hemodialysis) substantially impacts patient outcomes and payer cost. As of 2020, 480,516 ESRD patients were being treated with in-center hemodialysis, 65,406 with peritoneal dialysis, and 11,916 with home dialysis.[10] Studies show that peritoneal dialysis is associated with a higher quality of life than in-center hemodialysis, yet in-center hemodialysis remains the predominant form of treatment. Specifically, the all-cause fee-for-service (FFS) expenditures were $95,932 for hemodialysis patients in 2020, compared to $81,525 for peritoneal dialysis patients,[7] and cost savings associated with peritoneal dialysis have persisted, despite increasing adoption of peritoneal dialysis.[8] Moreover, additional studies have identified barriers to its adoption, including financial incentives that may favor in-center hemodialysis, and other patient-related factors such as health literacy, living conditions, and family support.[9] On the provider side, large dialysis organizations may be resistant to home dialysis given their investment in hemodialysis facilities and resources,[10] while workforce shortage and training deficiencies pose an obstacle for dialysis organizations of all sizes.[4]

[3]David Cutler et al., "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," American Economic Journal: Economic Policy 11, no. 1 (February 1, 2019): 192–221, https://doi.org/10.1257/pol.20150421.

[4] Ahmed, Salman, and Mallika L. Mendu. "The value of primary care provider involvement in the care of kidney failure patients on dialysis: Finding the balance." *Clinical Journal of the American Society of Nephrology* 15, no. 4 (2020): 450-452.

ESRD care requires multiple specialties and a high degree of care coordination, which is exacerbated by the complexity of the disease and its propensity to interact with other conditions.[5] Improving care for ESRD patients requires improved coordination across the care spectrum and shared responsibility among the multiple specialties involved for the patient's health outcomes.[6]

Furthermore, the ESRD patient population is at high risk for hospitalizations and associated adverse outcomes, including hospitalizations and readmissions. Research shows that patients on maintenance hemodialysis are at high risk for 30-day readmission, with over 70% of those readmissions being preventable. Additionally, catheter use for dialysis initiation remains problematic, with over 80% of incident hemodialysis patients initiating dialysis with a catheter between 2010 and 2020. Catheter use also increases risk of infections and impacts patient survival.

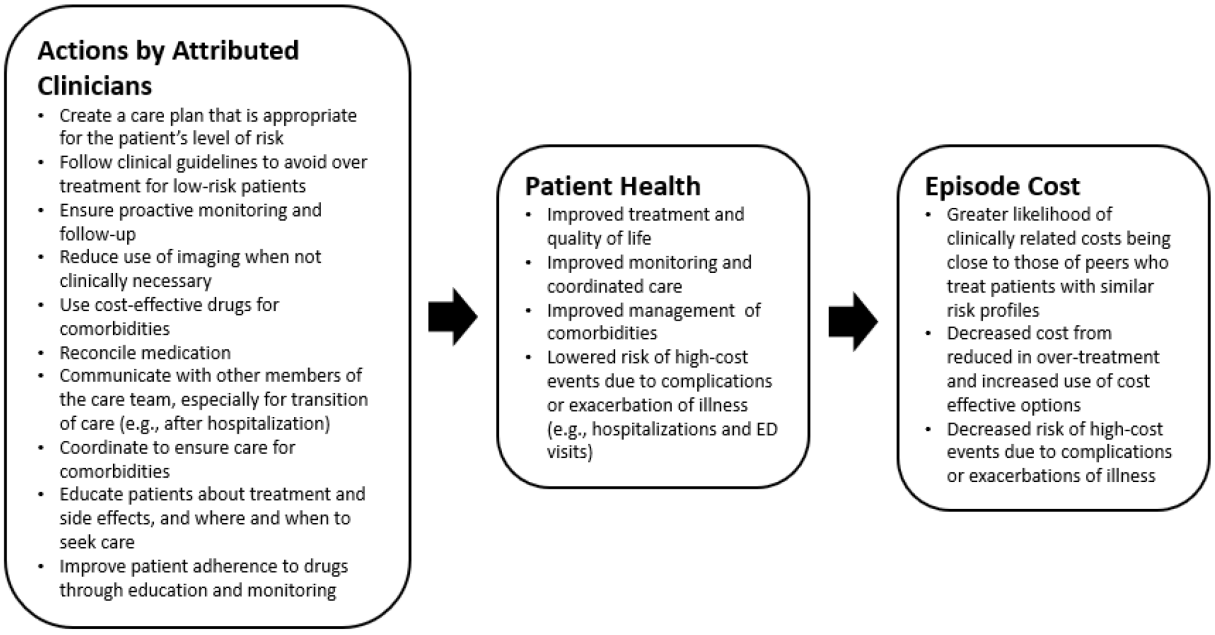
Managing ESRD poses a significant cost burden on healthcare systems, including Medicare. Patients with ESRD make up less than 1% of the Medicare population but contribute more than 7% of Medicare fee-for-service (FFS) payments.[7] The total inflation-adjusted Medicare expenditures for patients with ESRD increased steadily from $47.1 billion in 2010 to $53.0 billion in 2019.[10]

2.1.1 Logic Model

Figure 1: Logic Model of Steps between Actions by Attributed Clinicians and Episode Cost

[5] Marrufo G, Colligan EM, Negrusa B, Ullman D, Messana J, Shah A, Duvall T, Hirth RA. "Association of the Comprehensive End-Stage Renal Disease Care Model with Medicare Payments and Quality of Care for Beneficiaries with End-Stage Renal Disease," JAMA Intern Med. Case 180, no. 6 (2020): 852-860. DOI: 10.1001/jamainternmed.2020.0562.

[6] Sloan, Caroline E., Cynthia J. Coffman, Linda L. Sanders, Matthew L. Maciejewski, Shoou-Yih D. Lee, Richard A. Hirth, and Virginia Wang. "Trends in peritoneal dialysis use in the United States after Medicare payment reform." *Clinical Journal of the American Society of Nephrology* 14, no. 12 (2019): 1763-1772.



2.2 Performance Gap

2.2.1 Rationale

The ESRD episode-based measure was selected for development because of its high impact on Medicare spending, clinician coverage, and patient population, and it assesses costs for a chronic condition not captured by other measures in the MIPS cost performance category. The ESRD measure was also developed in consideration of alignment opportunities, notably the Kidney Care First (KCF) and Comprehensive Kidney Care Contracting (CKCC) payment Options of the Kidney Care Choices (KCC) Advanced Payment Model.

A measure-specific Clinician Expert Workgroup was then convened with clinicians, health care experts, and patient representatives who have appropriate experience to provide extensive, detailed input on this measure throughout its development.

2.2.2 Performance Scores

Table 1 shows the distribution of the measure score for clinician groups identified by a Tax Identification Number (TIN) and individual clinicians identified by a combination of a Tax Identification Number and National Provider Identifier (TIN-NPI).

Table 1 below shows substantial variations in cost performance for both TINs and TIN-NPIs, as evidenced by the interquartile ranges and score standard deviations. At the TIN and TIN-NPI levels, the 90th percentile score is much higher than the 10th percentile with over $11,000 and $13,000 difference in care costs for TINs and TIN-NPIs, respectively. Additionally, the difference between the minimum and maximum scores is $40,000 for TINs and $50,000 for TIN-NPIs. This suggests that there is an opportunity for improving clinician cost performance between the most and least efficient providers.

Table 1. Distribution of the Measure Score

Metric	TIN	TIN-NPI
Count	2,041	2,332
Mean Score	$63,578	$63,205

Metric	TIN	TIN-NPI
Score Standard Deviation	$4,685	$5,286
Minimum Score	$49,912	$48,112
Maximum Score	$99,537	$87,949
Score Interquartile Range (IQR)	$5,497	$6,983
Score Percentile		
10th	$58,130	$56,988
20th	$59,964	$58,777
30th	$61,224	$60,107
40th	$62,251	$61,429
50th	$63,282	$62,752
60th	$64,250	$64,159
70th	$65,488	$65,626
80th	$66,795	$67,411
90th	$69,479	$70,048

2.2.3 Disparities

Data on how the measure, as specified, addresses disparities is described in Sections 3.1.7 and 3.5.5.

3.0 Scientific Acceptability

3.1 Data Sample Description

Testing is based on the full population of measured entities and patients meeting inclusion and exclusion criteria for the measure, not based on a sample.

3.1.1 Type of Data Used for Testing

Medicare administrative claims data from the Common Working File (CWF), Long-Term Care Minimum Data Set (LTC MDS), and Medicare Enrollment Database (EDB).

3.1.2 Specific Dataset Used for Testing

The ESRD measure uses Medicare Part A, B and D claims data maintained by CMS. The cost measure uses Part A, B, and D claims to build episodes of care, calculate episode costs, and construct risk adjustors. These claims data are also used to designate episodes into clinically homogenous stratifications by sub-group and Part D enrollment status to ensure fair clinical comparisons among clinicians with a similar patient case mix Episode costs are payment standardized and risk adjusted to ensure accurate comparison of cost across clinicians. Payment standardization adjusts the allowed amount for a Medicare service to limit observed differences in costs to those that may result from healthcare delivery choices. Data from the EDB are used to determine beneficiary-level exclusions and secondary risk adjustors, specifically Medicare Parts A, B, and C enrollment, primary payer, disability status, end-stage renal disease (ESRD), patient birth dates, and patient death dates. The risk adjustment model also accounts for expected differences in payment for services provided to patients in long-term care based on data from the LTC MDS. Specifically, the LTC MDS is used to create the long-term care indicator variable in risk adjustment.

3.1.3 Dates of the Data Used in Testing

ESRD episodes ending from January 1, 2022, through December 31, 2022.

3.1.4 Levels of Analysis Tested

The measure was tested at group/practice (TIN) and individual clinician (TIN-NPI) levels.

3.1.5 Entities Included in the Testing and Analysis

Table 2 shows the individual clinician (identified by combination of TIN and NPI) and clinician group/practice (identified by TIN) included in the testing of the ESRD measure.

Table 2: Measured Entities Demographics

Metric	TIN		TIN-NPI	
	Count	%	Count	%
Count	2,041	100%	2,332	100%
Number of Episodes Attributed	-	-	-	-
20-39 Episodes	840	41.16%	1,983	85.03%
40-59 Episodes	407	19.94%	278	11.92%
60-79 Episodes	231	11.32%	55	2.36%
80-99 Episodes	139	6.81%	8	0.34%
100-199 Episodes	294	14.40%	6	0.26%
200-299 Episodes	80	3.92%	2	0.09%
300+ Episodes	50	2.45%	0	0%
Census Region	-	-	-	-

Metric	TIN		TIN-NPI	
No Data	Count	%	Count	%
Northeast	362	17.74%	289	12.39%
Midwest	398	19.50%	381	16.34%
South	879	43.07%	1,118	47.94%
West	390	16.24%	535	22.94%
Unknown	12	0.59%	9	0.39%

3.1.6 Patient Cohort Included in the Testing and Analysis

Table 3 shows the patient population for the ESRD measure testing. It consists of Medicare beneficiaries who meet measure criteria enrolled in Medicare Parts A and B whose chronic condition triggers an ESRD episode and do not meet the measure's exclusion criteria, as outlined in Section 3.4.1.

Table 3: Beneficiary Demographics

Metric	Value
Count	204,226
Mean Age	65.27 years
Female %	42.93%
Part D Enrollment %	80.35%

3.1.7 Social Risk Factors Included in Analysis

The analysis of social risk factors (SRFs) focused on examining the impact of Dual Medicare and Medicaid enrollment status on the measure. Table 4 outlines variables that may indicate SRFs and their advantages and disadvantages as indicators of individual-level SRFs. On balance, the analysis used dual Medicare and Medicaid enrollment status as the proxy of SRFs due to their broad availability in claims data, accurate measurement at the individual level, and wide acceptance of being a powerful indicator of health outcomes.[7]

Table 4: Social Risk Factors Available for Analysis

Variable	Advantages	Disadvantages	Used in Testing
Dual Medicare and Medicaid enrollment status	• Available for all beneficiaries • Most powerful predictor of poor outcomes[7]	• Variation in Medicaid eligibility across states	Yes
Race/Ethnicity	• Available for most beneficiaries, except for ambiguous categories of "Unknown" or "Other"	• Social risk driven by someone's race is often correlated with and partially captured by dual status[7]	No

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

Variable	Advantages	Disadvantages	Used in Testing
		• Only 5 categories available, which may lack granularity to fully capture disparities[8,9]	
ICD-10 Z codes for social determinants of health	• Reflects individual-level factors that influence health status and contact with health services	• Not routinely and consistently coded on claims, only available for 0.1% of all fee-for-service claims in 2019[10]	No
American Community Survey	• Can link beneficiary's zip code to socioeconomic (SES) measurement of their neighborhood • Many SES indices can be derived from the survey data (e.g., AHRQ index, deprivation index)	• Only a proxy measure, not always accurate at individual-level	No

3.2 Reliability Testing

3.2.1 Level of Reliability Testing

The following levels of reliability were tested: critical data elements used in the measure, group/practice (TIN) and individual clinician (TIN-NPI) levels.

3.2.2 Method of Reliability Testing

Data Element Reliability

The ESRD measure is constructed using CMS claims data, as described in Section 3.1.2. CMS has implemented several auditing programs to assess overall claims code accuracy, ensure appropriate billing, and recoup any overpayments.

- First, CMS routinely conducts data analyses to identify potential problem areas and detect fraud and audits necessary data fields used in this measure, including diagnosis and procedure codes and other elements consequential to payment. Specifically, CMS works with Zone Program Integrity Contractors, formerly Program Safeguard Contractors, to ensure program integrity; the agency also uses Recovery Audit Contractors to identify and correct for underpayments and overpayments.
- Second, CMS also uses the Comprehensive Error Rate Testing (CERT) Program to ensure that Medicare payments are correct under coverage, coding, and billing rules.

[8] Nguyen, Kevin H., Kaitlyn P. Lew, and Amal N. Trivedi. "Trends in Collection of Disaggregated Asian American, Native Hawaiian, and Pacific Islander Data: Opportunities in Federal Health Surveys." *American Journal of Public Health* (2022).
[9] Kader, Farah, Lan N. Doan, Matthew Lee, Matthew K. Chin, Simona C. Kwon, and Stella S. Yi. "Disaggregating Race/Ethnicity Data Categories: Criticisms, Dangers, And Opposing Viewpoints", *Health Affairs Forefront* (2022).
[10] Centers for Medicare and Medicaid, Office of Minority Health. "Utilization of Z Codes for Social Determinants of Health among Medicare Fee-for-Service Beneficiaries." (2019) https://www.cms.gov/files/document/z-codes-data-highlight.pdf

CMS continues to perform corrective actions and give providers additional education to ensure accurate billing.

- Lastly, to ensure claims completeness and inclusion of any corrections, the measure was developed and tested using data with three-month claims run-out from the end of the measurement period.

Clinician-level Reliability

Measure reliability is the degree to which repeated measurements of the same entity agree with each other). For measures of clinician performance, the measured entity is the TIN or TIN-NPI, and reliability is the extent to which repeated measurements of the TIN or TIN-NPI give similar results. To estimate measure reliability, we use a signal-to-noise analysis.

This approach seeks to determine how much of the variation in the measure score is explained by differences among clinician performance (i.e., signal) rather than random variation (i.e., statistical noise) among clinicians due to the sample of cases observed. To achieve this, we calculate reliability scores as:

$$R_j = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_{w_j}^2}$$

Where:

$\sigma_{w_j}^2$ is the within-group variance of the mean measure score of clinician *j*

σ_b^2 is the between-group variance of clinicians within the episode group

That is, reliability is calculated as the ratio of between-group variance to the sum of between-group variance and within-group variance. Reliability closer to a value of 1.0 indicates that the between-group variance is relatively large compared to the within-group variance, which suggests that the measure is effectively capturing the systematic differences between the clinician and their peer cohort.

3.2.3 Statistical Results from Reliability Testing

Data Element Reliability

Between 2005 and 2020, CMS Comprehensive Error Rate Testing (CERT) estimates that proper payment, which includes payments that met Medicare coverage, coding, and billing rules, ranged from 87.3% to 93.7% of total payments each year.[11] The fiscal year 2022 Medicare fee-for-service program proper payment rate was 92.5%.[12]

Clinician-level Reliability

The table below shows reliability metrics at the 20-episode testing volume threshold. 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 testing purposes, we used a 20-episode volume threshold. If the measure is implemented in MIPS in the future, CMS will establish a case minimum through notice-and-comment rulemaking.

[11]Comprehensive Error Rate Testing (CERT) Program. "Appendices Medicare Fee-for-Service 2020 Improper Payments Report". Table A6. https://www.cms.gov/files/document/2020-medicare-fee-service-supplemental-improper-payment-data.pdf-1.
[12]Ibid.

Table 5: Reliability at the Accountability Entity Level

Reporting Level	Entities Meeting Case Minimum	Mean Reliability	Median Reliability	% Above 0.4	% Above 0.7
TIN	2,041	0.52	0.51	69.87%	18.77%
TIN-NPI	2,332	0.37	0.29	39.71%	0.90%

3.2.4 Interpretation

The results of the data element testing show moderate reliability of the critical data elements used by the measure. At the entity level, the measure is moderately reliable at the TIN reporting level (0.52) and 69.87% of TINs meet or exceed the moderate reliability threshold of 0.4. At the TIN-NPI level, the mean reliability of the measure is 0.37 and 39.71% of TIN-NPIs meet or exceed this moderate reliability threshold. Reliability is one way to consider the extent to which performance comparisons among clinicians reflect systematic differences in performance. CMS considered existing scientific evidence on various interpretations and methods of estimating reliability. In the CY 2022 Physician Fee Schedule (86 FR 64996) rule, CMS reaffirmed the 0.4 threshold for mean reliability, continues to be appropriate for indicating moderate reliability for performance measures in the Cost category of the MIPS program.[13]

3.3 Validity Testing

3.3.1 Level of Validity Testing

The validity of the measure was tested using empirical validity at the accountable entity level (TIN and TIN-NPI).

3.3.2 Method of Validity Testing

Face Validity

The ESRD measure was developed through a structured, iterative process for gathering detailed input on the measure from recognized clinician experts. Experts in this clinical area evaluated specifications to ensure that each aspect of the measure (e.g., assigned services) was intentionally capturing only the costs of care within the reasonable influence of the attributed clinician for a defined patient population (i.e., the ability of the measure score to differentiate between good from poor performance).

In developing this measure, Acumen incorporated input from:

(i) a Chronic Kidney Disease/ESRD Clinician Expert Workgroup;
(ii) a Technical Expert Panel (TEP);
(iii) the Person and Family Partners; and
(iv) a national field testing period.

This process is detailed in the Episode-Based Cost Measures Development Process document posted on the QPP Cost Measure Information Page.[14]

One of the primary roles of the Clinician Expert Workgroup is to develop service assignment rules for the cost measure. These service assignment rules seek to ensure clinicians are

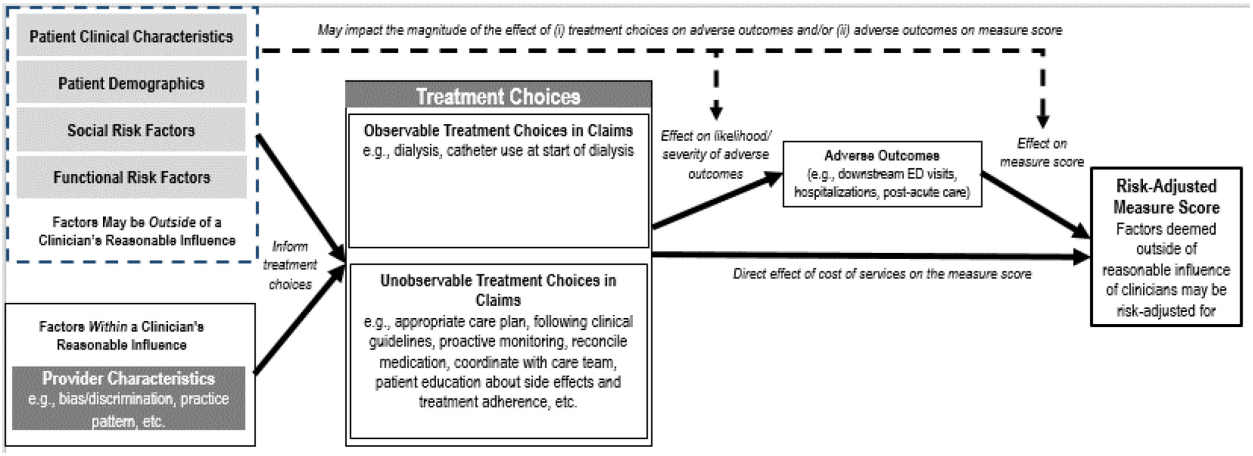
[13] CMS, "Medicare Program; CY 2022 Payment Policies Under the Physician Fee Schedule and Other Changes to Part B Payment Policies; Medicare Shared Savings Program Requirements; Provider Enrollment Regulation Updates; and Provider and Supplier Prepayment and Post-Payment Medical Review Requirements,"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

evaluated on services and costs that are clinically related to the attributed clinician's role in managing ESRD, thus limiting cost variation unrelated to clinician care in this measure. Therefore, assigned services are services that the Clinical Expert Workgroup believed an attributed clinician could influence their occurrence, frequency, or intensity.

Empirical Validity Testing

Validity is a criterion used to assess whether the cost measure can quantify the construct it aims to measure, which is the cost directly related to treatment choices and the cost of adverse outcomes resulting from care. We evaluated the empirical validity of the ESRD measure by estimating the effect of relevant treatment choices on the measure score using multiple regression, based on the conceptual model outlined in Figure 2.

Figure 2: Conceptual Model of Treatment Choices on the Measure Score



The cost measure is designed to reflect costs directly related to treatment choices, and the cost of adverse outcomes resulting from care. Therefore, treatment choices, either observable in claims or otherwise, by an attributed clinician can directly impact the measure score or indirectly when they are mediated through the cost of adverse outcomes. In turn, the cost of adverse effects to the total cost captured by the measure score.

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

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

3.3.3 Statistical Results from Validity Testing

Empirical Validity Testing

Table 6 shows two regression models for each reporting level. Model 1 shows the effect on the clinicians' mean observed cost to expected cost ratio for each additional one thousand dollar of a cost category that is assigned to an episode, on average, while holding the remaining

categories of cost constant. Model 2 shows the effect on the mean cost of adverse events for each additional one thousand dollar of a cost category that is assigned to an episode, on average, while holding the remaining categories of cost constant.

Table 6. Estimated Effect on Treatment Choices on the Measure Score

Service Categories	Coefficient in Thousands [95% Confidence Interval] (p-value)			
	TIN		TIN-NPI	
	Model 1: Mean O/E = Mean Cost of Treatment Choices + Mean Cost of Adverse Events	Model 2: Mean Cost of Adverse Events = Mean Cost of Treatment Choices	Model 1: Mean O/E = Mean Cost of Treatment Choices + Mean Cost of Adverse Events	Model 2: Mean Cost of Adverse Events = Mean Cost of Treatment Choices
Adverse Events	0.01 [0.01,0.01] (p < 0.01)	-	0.01 [0.01,0.01] (p < 0.01)	-
Outpatient Evaluation & Management (E/M) Services	-0.01 [-0.01,0.00] (p < 0.01)	3.68 [3.19,4.17] (p < 0.01)	0.0 [0.00,0.01] 1.0 (p = 0.32)	2.38 [1.87,2.89] (p < 0.01)
Major Procedures	0.01 [0.01,0.02] (p < 0.01)	-0.64 [-1.22,-0.05] (p = 0.03)	0.01 [0.01,0.01] (p < 0.01)	-0.12 [-0.56,0.32] (p = 0.61)
Ambulatory/Minor Procedures	0.02 [0.01,0.02] (p < 0.01)	0.89 [0.20,1.59] (p = 0.01)	0.01 [0.00,0.01] (p < 0.01)	0.70 [0.18,1.22] (p < 0.01)
Laboratory, Pathology, and Other Tests	0.02 [0.00,0.04] (p = 0.06)	-3.88 [-6.00,-1.76] (p < 0.01)	-0.01 [0.03,0.01] (p = 0.44)	-0.08 [-1.86,1.70] (p = 0.93)
Imaging Services	0.02 [0.00,0.03] (p = 0.03)	-0.17 [-1.69,1.36] (p = 0.83)	0.00 [-0.01,0.02] (p = 0.51)	0.68 [-0.62,1.98] (p = 0.31)
Durable Medical Equipment (DME)	0.0 [0.00,0.01] (p = 0.31)	1.03 [0.29,1.78] (p < 0.01)	0.01 [0.00,0.02] (p = 0.13)	0.90 [0.04,1.75] (p = 0.04)
Anesthesia Services	0.00 [-0.03,0.02] (p = 0.73)	14.26[11.27,17.24] (p < 0.01)	0.02 [-0.01,0.05] (p = 0.15)	11.37 [8.82,13.92] (p < 0.01)
Chemotherapy and Other Part-B Covered Drugs	0.01 [0.01,0.01] (p < 0.01)	-0.03 [-0.24,0.18] (p = 0.79)	0.01 [0.01,0.01] (p < 0.01)	-0.07 [-0.28,0.14] (p = 0.51)
Part-D Drugs	0.01 [0.01,0.01] (p < 0.01)	0.08 [-0.12,0.27] (p = 0.44)	0.01 [0.01,0.01] (p < 0.01)	0.20 [0.05,0.35] (p < 0.01)
Dialysis	0.01 [0.01,0.01] (p < 0.01)	-0.66 [-0.78,-0.53](p < 0.01)	0.01 [0.01,0.01] (p < 0.01)	-0.40 [-0.49,-0.31] (p < 0.01)
All Other Services Not Otherwise Classified	0.04 [-0.02,0.11] (p = 0.16)	2.89 [-3.85,9.64] (p = 0.40)	0.04 [0.00,0.08] (p = 0.07)	-0.83 [-4.69,3.04] (p = 0.68)

3.3.4 Interpretation

Overall, testing results demonstrate that the cost measure reflects the cost directly related to treatment choices and the cost of related adverse outcomes. Therefore, there is evidence that the measure captures what it purports to measure.

Model 1 shows that adverse events is associated with worse measure score. Dialysis and Part B medications are associated with worse measure score but they are shown to be associated with lower cost of adverse outcomes, which suggests that they are important to patient outcomes and are likely not good candidates for cost reduction. Major procedures also show similar pattern. Minor procedures, and Part D medications are associated with both worse measure score and adverse events, which suggest a potential overuse or frequent co-occurrence with adverse events.

3.4 Exclusions Analysis

3.4.1 Method of Testing Exclusions

Exclusions are used in the ESRD measure to ensure a comparable patient population within the scope of the measure's focus on ESRD management and that episodes provide meaningful information to attributed clinicians. Exclusions are also used as part of data processing so that sufficient data are available to accurately determine episode spending and calculate risk adjustment for each episode.

For the exclusions analysis discussed in this section, we focused on exclusion criteria intended to ensure a comparable patient population.

- Episodes where patient death date occurred before the episode end date
    - These episodes were excluded as they may not accurately reflect a clinician's performance as the truncated episode window does not capture the full length of care intended by the measure.
- Episodes that ended in transplant
    - These episodes were excluded as they are often clinically distinct from the overall end-stage renal disease population.

Given the rationales for these exclusions, we expect these excluded episodes to have a different profile than the included episodes, such as a higher mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). For each exclusion, we examined the number of episodes and beneficiaries affected, as well as the distributions of observed cost. We then compared the cost characteristics of the excluded episodes to those of episodes included in the measure calculation to assess the distinctness between the two patient cohorts. A full list of the exclusions used for the ESRD measure is provided in the Measure Codes List available on the QPP Cost Measure Information Page.[15]

3.4.2 Statistical Results from Testing Exclusions

Table 7 below presents 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.

[15]CMS, QPP Cost Measure Information Page, https://www.cms.gov/medicare/quality/value-based-programs/cost-measures.

Table 7: Cost Statistics for Measure Exclusions

Exclusion	Episodes		Observed Cost					
				Percentile				
	#	% of All Episodes Meeting Triggering Logic	Mean	10th	25th	50th	75th	90th
All Episodes Meeting Triggering Logic	294,679	100%	$74,235	$43,392	$48,963	$59,481	$82,450	$122,119
Episode Length Less Than One Attribution Window	32,034	10.87%	$122,585	$44,359	$56,541	$86,902	$145,674	$238,580
Beneficiary Death in Episode	80,357	27.27%	$102,991	$45,596	$56,670	$79,303	$119,108	$181,112
Outlier	4,084	1.39%	$81,304	$2,965	$7,384	$93,509	$154,885	$154,885
TIN does not Meet Case Minimum	61,325	20.81%	$80,545	$43,814	$49,751	$62,300	$89,851	$137,207
No Attributed NPI	37,429	12.70%	$76,941	$44,603	$50,610	$62,419	$86,935	$126,911
TIN-NPI does not Meet Case Minimum	163,652	55.54%	$76,209	$43,405	$49,051	$59,954	$84,428	$127,044
Episode Ended in Transplant	10,096	3.43%	$59,740	$44,304	$48,200	$53,986	$64,636	$79,929
Reportable Episodes (if all clinicians reported as TIN at the Testing Volume Threshold)	160,669	54.52%	$62,662	$43,108	$47,874	$55,491	$70,877	$94,593
Reportable Episodes (if all clinicians reported as TIN-NPI at the Testing Volume Threshold)	67,185	22.80%	$61,523	$42,748	$47,465	$54,776	$69,373	$91,912

3.4.3 Interpretation

Table 7 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 highlighting that only the observed cost is shown, which has not been risk adjusted. Therefore, the differences in cost may appear much smaller after risk adjustment than as-is. Overall, the exclusion criteria decrease the distribution of observed cost on the measure shown by the large disparities between the reportable episode observed cost and the excluded episode observed costs.

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

Episodes where a beneficiary dies before the episode end date are excluded because they do not provide sufficient data in the episode window period. These episodes also have a higher mean observed cost than all episodes meeting triggering logic, at $102,991, likely because the costs are distributed over fewer days than a typical episode.

Episodes classified as outlier cases are excluded because they deviate substantially from the projected cost for a given patient risk profile. Outlier episodes have a mean observed episode cost of $81,304 compared to $66,422 for all episodes meeting triggering logic. The wide variability of observed episode costs for outlier cases also supports their exclusion. At the 10th percentile the outlier cases observed cost is $2,965, and at the 90th percentile the observed cost is $154,885.

Episodes that end in transplants are excluded because they are often clinically distinct from the overall end-stage renal disease population; they are shorter than an average episode and have lower observed cost. The decision to exclude these episodes is based on input from the clinical expert workgroup during the measure development process.

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

3.5 Risk Adjustment or Stratification

3.5.1 Method of Controlling for Differences

Differences in case mix are controlled for using a statistical risk model with 110 risk factors and stratification by 2 risk categories.

The risk adjustment model for the ESRD measure adjusts for comorbidities based on the CMS Hierarchical Condition Category (HCC) model, count of HCCs, end-stage renal disease (ESRD) status, disability status, number and types of clinician specialties from which the patient has received care, recent use of institutional long-term care, age, and dual eligibility status.

The model also includes measure-specific factors:

- Crash starts to dialysis (unplanned transition to ESRD)

A separate linear regression is run for each sub-group and Medicare Part D enrollment status combination to ensure fair comparison:

The episode's scaled (i.e., annualized) observed costs are winsorized at the 98th percentile prior to the regression for each model to handle extreme observations. Full details of the risk adjustment model are in the Measure Codes List File available on the QPP Cost Measure Information page.[16]

3.5.2 Conceptual, Clinical, and Statistical Methods

We selected the CMS-HCC model based on previous studies evaluating its appropriateness for use in risk adjusting Medicare claims data. This model was developed specifically for use in the Medicare population, meaning that it accounts for conditions found in the Medicare population. In addition, the CMS-HCC model is routinely updated for changes in coding practices (e.g., the

[16]CMS, QPP Cost Measure Information Page, https://www.cms.gov/medicare/quality/value-based-programs/cost-measures.

transition from ICD-9 to ICD-10 codes). Because the CMS-HCC model has already been extensively tested, we focus our testing on the adaptation of the CMS-HCC model to the ESRD measure's patient population.

The workgroup provided input on measure-specific risk adjustors after reviewing empirical analyses on subpopulations of interest to assess whether and if so, how, particular factors should be accounted for in the model. These could include patient characteristics, factors outside of the reasonable influence of the clinician, or any other factors that would help prevent unintended consequences. These additional risk adjustors are listed in the section above.

As previously noted, the risk adjustment model is run on episodes stratified into episode sub-groups, which may qualify as "ordering" of risk factors. Episode sub-groups were also determined based on the workgroup's input, with the goal of ensuring clinical comparability among episodes so that the cost measure fairly compares clinicians with similar patient case-mix.

3.5.3 Conceptual Model of Impact of Social Risks

Figure 2 shows the conceptual model that outlines how SRFs can influence the measure score, which is informed by published external research and Acumen's data analysis.[7,17,18,19,20] The conceptual model outlines risk factors that are either known by the literature or informed by the Clinical Expert Workgroup to be within or outside the influence of the attributed clinician. Risk factors, including SRFs, can influence the treatment choices and impact the size of the effect of treatment choices on mitigating the risk and cost of adverse outcomes.

A systematic approach then guides the decision of which factors to include in the risk adjustment model:

1. 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 HCCs.
2. 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 patient characteristics. We arrived at the final list of risk adjustors based on those discussions and consensus among the clinical experts.
3. During our testing phases, we also follow a structured and systematic approach to deciding whether SRFs should be adjusted for, further described in Section 3.5.5.

3.5.4 Statistical Results

The literature has extensively tested using the HCC model for Medicare claims data. Although the variables in the HCC model were selected to predict annual cost, CMS has also used this

[17]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.
[18]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
[19]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/.
[20] 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-medicares-value-basedpurchasing-programs

risk adjustment model in several other settings (e.g., Accountable Care Organizations, previous physician Quality and Resource Use Report programs, and other administrative claims-based measures such as the Knee Arthroplasty episode-based cost measure, Total Per Capita Cost (TPCC) cost measure, Medicare Spending Per Beneficiary (MSPB)-PAC cost measure and MSPB-Hospital cost measure). Recalling that the risk model relies on the existing CMS-HCC model, testing results for factors included in the CMS-HCC V24 model can be found in the Evaluation of the CMS-HCC Risk-Adjustment Model report[21] and the Report to Congress: Risk Adjustment in Medicare Advantage[22]. For measure-specific factors not included in the CMS-HCC model, we sought expert clinician input through the workgroup, which provided recommendations on additional risk adjustors and sub-groups.

3.5.5 Analyses and Interpretation in Selection of Social Risk Factors

To determine whether it is appropriate to risk adjust for SRFs, the following criteria are considered:

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

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

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

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

Table 8: Coefficient of Patient-level Dual Status under Different Models

Level	% of All Episodes	Coefficient of Patient-level Dual Status (P-value)		
		Base Model + Patient-level Dual Status	Base Model + Patient-level Dual Status + Clinician's Dual Share	Base Model + Patient-level Dual Status + Clinician's Fixed Effect
TIN w/ Part D	55.69%	0.03 (p:<0.0001)	0.02 (p:<0.0001)	0.02 (p:<0.0001)
TIN w/out Part D	1.70%	0.04 (p:0.15)	0.04 (p:0.14)	0.01 (p:0.72)
TIN-NPI w/Part D	56.03%	0.02 (p:<0.0001)	0.02 (p:<0.0001)	0.03 (p:<0.0001)
TIN-NPI w/out Part D	1.67%	0.02 (p:0.35)	0.02 (p:0.37)	0.01 (p:0.69)

[21]Pope, Gregory C., John Kautter, et al., "Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report." RTI International: March 2011.

[22]CMS, "Report to Congress: Risk Adjustment in Medicare Advantage," https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/RTC-Dec2018.pdf.

Table 9: Mean Ratio of Episode Observed Cost to Expected Cost (O/E) Stratified by Clinician's Dual Share and Patient's Dual Status

Dual Share	TIN			TIN-NPI		
	All Episodes	Dual Episodes	Non-Dual Episodes	All Episodes	Dual Episodes	Non-Dual Episodes
(ALL)	1.00	1.01	1.00	1.00	1.00	1.00
0%-20%	1.00	1.01	1.00	1.00	1.01	0.99
21%-40%	1.00	1.01	1.00	1.00	1.00	1.00
41%-60%	1.00	1.00	0.99	1.00	1.00	1.00
61%-80%	1.01	1.01	1.00	1.01	1.01	1.00
81%-100%	1.01	1.01	1.02	1.00	1.00	1.00

Table 10. Proportions of Clinicians Who Perform Significantly Worst, Equally Well, or Significantly Better on Their Dual Episodes than Non-Dual Episodes

Reporting Level	Significantly Worse	Equally Well	Significantly Better
TIN	3.98%	93.48%	44.83%
TIN-NPI	2.66%	94.33%	45.64%

Table 11. Clinicians' Performance Shift after Adding a Dual Status Risk Adjustor

TIN or TIN-NPI	Proportion of Clinicians Affected at Various Levels of Performance Shift	
	Ranking Shift by 1% or more	Ranking Shift by 5% or more
TIN	65.75%	4.02%
TIN-NPI	61.47%	2.23%

There's a statistically significant association between the patient's dual status and episode cost for episodes with Part D coverage, which is the largest subgroup (Table 8). This association is relatively stable in the largest subgroups and maintains statistically significance even after adding variables to account for clinician-level factors, which suggests that the patient-level factors are more influential than clinician-level factors. This is also supported by the evidence that the performance degradation is observed mainly on dual episodes (Table 9). The majority of clinicians are able to perform equally well or significantly better on their dual episodes and non-dual episodes, while a small percentage perform significantly worse on their dual episodes than their non-dual episodes, which suggests that clinicians are able to fully mitigate the effect of SRFs (Table 10). Lastly, risk adjusting for dual status appears to change the performance ranking for many clinicians (Table 11).

3.5.6 Method for Statistical Model or Stratification Development

To analyze the validity of current risk adjustment model, we examined two criteria: discrimination and calibration.

1) 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. These results are provided in Section 3.5.7.

2) 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. A well-calibrated measure should have predictive ratios close to 1.0 across all deciles. These are discussed in Sections 3.5.8 and 3.5.9.

3.5.7 Statistical Risk Model Discrimination Statistics

The overall R-squared for the ESRD cost measure, calculated by dividing explained sum of squares by total sum of squares is 0.171. The adjusted R-squared is 0.170. More information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.[23]

3.5.8 Statistical Risk Model Calibration Statistics

The predictive ratio is calculated using the formula of average expected cost / average observed cost for all episodes in each decile.

3.5.9 Statistical Risk Model Calibration – Risk Decile

Analysis of predictive ratios by risk decile for the measure shows moderate variation among risk deciles, as predictive ratios range from 0.93 to 1.02 across all risk deciles (with an overall average of 1).

Table 12: Predictive Ratio by Decile of Predicted Episode Cost

Decile	Average Predictive Ratio
Decile 1	0.93
Decile 2	1.00
Decile 3	1.01
Decile 4	1.01
Decile 5	1.02
Decile 6	1.02
Decile 7	1.02
Decile 8	1.00
Decile 9	0.99
Decile 10	0.99

3.5.10 Interpretation

The R-squared values for the model, which measure the percentage of variation in results predicted by the model, are higher than the values presented in similar analyses of risk adjustment models.[24] As noted in Section 3.5.6 and 3.5.7, these results should be interpreted alongside service assignment rules, which remove clinically unrelated services.

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

[23]Pope, Gregory C., John Kautter, et al., "Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report." RTI International: March 2011.
[24]Pope, Gregory C., John Kautter, Melvin J. Ingber, Sara Freeman, Rishi Sekar, and Cordon Newhart. "Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report." RTI International: March 2011.

and differentiate the performance of clinicians. Therefore, achieving high explained variance is optional because the measure should only adjust for some variations in the cost of care. In collaboration with the experts from our clinical workgroup, this measure only adjusts for factors that are deemed outside the reasonable influence of clinicians. The service assignment rules provide context for which costs are included in the measure and which are not.

Table 12 shows that the risk adjustment model is moderately consistent, with the average predictive ratios observed to be close to 1.00 across all deciles, with the range between 0.93 and 1.02.

3.6 Identification of Meaningful Differences in Performance

3.6.1 Method

To identify meaningful differences in performance, this analysis first examines the distribution of the measure score to highlight the performance gap between the most and least efficient clinicians. Then, this analysis examines the rate of adverse events that may occur during an episode of care to highlight the variation in frequency and cost of those events.

3.6.2 Statistical Results

Table 1 shows the distribution of the measure score at the TIN and TIN-NPI levels. There is a difference in mean score for TIN and TIN-NPI levels because each level has its own attribution rules, which resulted in slightly different populations of episodes used for measure score calculation. However, clinicians are only compared to their peers at either the TIN or TIN-NPI level, therefore the differences in score across different levels can be ignored.

The rate of inpatient readmissions stays is observed to be at 35.5% with an average episode cost of $84,343. The rate of emergency services is observed at 60.39% with an average cost of $69,562.

3.6.3 Interpretation

There are substantial variation observed in the measure score in both TIN and TIN-NPI levels, indicated by the interquartile ranges, standard deviations, and coefficients of variation. The magnitude of the observed variation is in the thousands of dollars, which indicates that there are opportunities to close the gaps between the most and least efficient clinicians.

Since each episode with readmissions and emergency department visits is very costly, every percentage reduction in readmission and emergency department visit rates represents substantial performance improvement for the attributed clinician or clinician group.

3.7 Missing Data Analysis and Minimizing Bias

3.7.1 Method

Since CMS uses Medicare claims data to calculate the ESRD measure, Acumen expects a high degree of data completeness. To further ensure that we have complete and accurate data for each patient, Acumen excludes episodes where patient date of birth information (an input to the risk adjustment model) cannot be found in the EDB, the patient does not appear in the EDB, or the patient death date occurs before the episode trigger date.

The ESRD measure also excludes episodes where the patient is enrolled in Medicare Part C or has a primary payer other than Medicare in the 120-day lookback period and episode window. In such situations, Medicare Parts A and B claims data may not capture the complete clinical profile for the patient needed to capture the clinical risk of the patient in risk adjustment. Furthermore, Parts A and B claims data may not capture all Medicare resource use if some portion of the patient's care is covered under Medicare Part C.

3.7.2 Missing Data Analysis

The table below presents the frequency of missing data across the categories of missing data which caused episodes to be excluded from the ESRD measure. Frequency is presented in terms of the number of episodes excluded due to missing data, as well as the cost profile of episodes with missing data compared to episodes included in the measure reporting.

As a note, the episode and clinician counts below reflect exclusion from the initial population of triggered episodes. After the missing data exclusions are applied, we apply additional exclusions, as outlined in section 3.4, to this overall patient cohort to narrow the population to only applicable episodes.

Table 13: Cost Statistics for Missing Data Category

Missing Data Categories	Episodes	Observed Cost					
	#	Mean	Percentile				
			10th	25th	50th	75th	90th
All Episodes	499,549	$66,422	$16,824	$44,551	$55,302	$77,082	$115,721
Beneficiary Resides Outside of U.S. or Territories	10,305	$49,109	$9,420	$22,443	$46,506	$60,705	$85,091
No Continuous Enrollment in Medicare Parts A and B, and Any Enrollment in Part C	117,695	$39,853	$3,465	$9,408	$30,587	$53,863	$80,665
Primary Payer Other than Medicare	61,425	$58,422	$4,614	$27,290	$51,217	$71,129	$107,974

3.7.3 Interpretation

The results show that the missing data episodes appear to be different than all episodes in terms of observed episode cost (Table 13). The number of episodes and variability in observed costs suggest these excluded episodes could impact the measure if they were included. However, these episodes are likely to have insufficient data to be comparable to all other episodes, thus making it appropriate to exclude these episodes from the measure.

4.0 Feasibility

4.1 Data Elements Generated as Byproduct of Care Processes

The data elements used in this measure are pulled from Medicare claims. They can be based on information generated, collected and/or used by healthcare personnel during the provision of care (e.g., diagnoses), which are then translated into the appropriate coding system (e.g. ICD-10 diagnoses, MS-DRGs) for use in Medicare claims by either the original healthcare personnel or another individual.

4.2 Electronic Sources

All data elements are in defined fields in electronic claims.

4.3 Data Collection Strategy

4.3.1 Data Collection Strategy Difficulties

Lessons and associated modifications may be categorized into three types: data collection procedures, handling of missing data, and sampling data associated with beneficiaries who died during an episode of care.

4.3.1.1 Data Collection

Acumen receives claims data directly from the CWF maintained at the CMS Baltimore Data Center. Healthcare providers submit Medicare claims to a Medicare Administrative Contractor (MAC), which are subsequently added to the CWF. However, these claims may be denied or disputed by the MAC, leading to changes to historical CWF data. In rare circumstances, finalizing claims may take many months or even years. As such, it is not practical to wait until all claims for a given month are finalized before calculating the measure, resulting in a trade-off between efficiency (accessing the data on time) and accuracy (waiting until most claims are finalized) when determining the duration (i.e., the “claims run-out” period) after which to pull claims data. To determine the appropriate claims run-out period, Acumen has tested the delay between claim service dates and claims data finalization. Based on this analysis, Acumen uses a run-out period of three months after the end of the calendar year to collect data for development and testing purposes. If CMS adopts this measure for use in a program, calculation and reporting would align with the program’s reporting practices.

4.3.1.2 Missing Data

This measure requires complete beneficiary information, therefore, a small number of episodes with missing data are excluded to ensure data completeness and accurate comparability across episodes. For example, episodes where the beneficiary was not enrolled in Medicare Parts A and B for the 120 days before the episode start date are excluded from this measure. Excluding these episodes enables the risk adjustment model to accurately adjust for the beneficiary’s comorbidities using data from the previous 120 days of Medicare claims. Additionally, the risk adjustment model includes a categorical variable for beneficiary age bracket, so episodes for which the beneficiary’s date of birth cannot be located are excluded from the measure.

4.3.1.3 Sampling

During measure testing, Acumen noted that episodes in which the beneficiary died before the episode end date exhibited different cost distributions than other episodes. As such, this measure excludes episodes to avoid negatively impacting clinician scores.

5.0 Usability and Use

5.1 Use

5.1.1 Current and Planned Use

The measure is not currently in use but is intended for use in a payment program and could eventually be publicly reported. It was specifically developed for potential use in the Cost performance category of MIPS to assess clinicians reporting as individuals or groups under a contract with CMS.

For CMS to approve this measure for use in MIPS, it must be reviewed by the Pre-Rulemaking Measure Review process (PRMR; formerly referred to as the Measure Application Partnership [MAP]) and then undergo the notice-and-comment process. Given these next steps, the earliest the measure could be used in MIPS is CY 2025. If in use, CMS can then determine whether to publicly report the cost measure.

5.1.2 Feedback on the Measure by Those being Measured or Others

Throughout the ESRD measure development, we used an iterative and extensive process to gather feedback on the measure and its results to ensure that it can be used appropriately in the MIPS program by clinicians and clinician groups who practice in this clinical area. This process also seeks to ensure that the measured entities can understand and interpret their performance results to help support decision-making. A couple of the main ways we gathered input was through reoccurring Clinician Expert Workgroup meetings, which incorporated feedback from the patient and caregiver perspective, empirical data, and discussion between clinician experts who recommend measure specifications, and through the national field testing of the measures.

5.1.2.1 Technical Assistance Provided During Development or Implementation

Clinician Expert Workgroup Meetings

For each Clinician Expert Workgroup meeting, Acumen provided empirical data (e.g., analyses on potentially relevant services to group and potential sub-populations to sub-group, risk adjust, or exclude) to inform the Clinician Expert Workgroup members' recommendations. These analyses were conducted using all administrative claims data for Medicare Parts A, B, and D. This data was shared with Workgroup members to help inform their feedback on the measure specifications throughout its development to ensure that the measure is appropriately assessing costs for these clinicians.

Field Testing

Additionally, Acumen and CMS nationally field tested the draft ESRD measure, along with 4 other episode-based cost measures, for a 4-week comment period (January 17 to February 14, 2023). We provided a Field Test Report with performance data to all clinician groups and clinicians who were attributed 20 or more episodes, which was the testing volume threshold.[25] This testing sample was selected to balance coverage and reliability, since a key goal of field testing was to test the measures with as many stakeholders as possible. A total of 5,369 reports were developed for this measure. During this time, feedback was gathered on the usability of the performance data and the appropriateness of the measure.

[25]The field test reports were available for download from the Quality Payment Program website: https://qpp.cms.gov/login.

5.1.2.2 Technical Assistance with Results

Clinician Expert Workgroup Meetings

Acumen provided data before or during each of the Clinician Expert Workgroup Meetings: The Workgroup Webinar, Service Assignment and Refinement Webinar, and Post-Field Test Refinement Webinar. During the meetings, Acumen would guide Workgroup members through these analyses, providing clinical and programmatic context when needed. Using this iterative process, the Workgroup members discussed the testing results in depth during each meeting and allowed the data to inform their recommendations for measure specifications. The goal was to ensure that the measure appropriately assessed clinicians' cost of care within their reasonable influence without creating potential unintended consequences so that it could be usable in the MIPS program.

Field Testing

During the field testing period, the measured entities (i.e., MIPS-eligible clinicians and clinician groups who received a report) and the general public provided feedback on the appropriateness of the measures and the usability of the data. The public comments were summarized in a report, which was shared with the Clinician Expert Workgroup for consideration when recommending refinements to the measures based on the testing data and feedback.

The following sections offer more details on the contents of each report and describe the education and outreach efforts associated with the field testing feedback period.

Data Provided During Field Testing

Each Field Test Report contained:

- Detailed performance results for the attributed measure, including cost measure score and breakdown of episode cost compared to the national average and TIN/TIN-NPIs with a similar patient case mix (or risk profile).
- Drill-down detail for each measure, including more detailed information on potential cost drivers in the TIN/TIN-NPI's episodes. For example:
    - Analysis of utilization and cost for the measure by the Restructured BETOS Classification System (e.g., outpatient evaluation and management services, procedures, and therapy, hospital inpatient services, emergency room services, post-acute services)[26]
    - Breakdown of costs for Part B Physician/Supplier and inpatient claims (e.g., top 5 most billed services and by risk bracket)
    - Accompanying episode-level Comma Separated Value (CSV) file with detailed information for all episodes attributed to the TIN/TIN-NPI. This file provides detailed information on every episode used to calculate your measure score, which includes winsorized observed cost, risk-adjusted cost, facilities and clinicians rendering care, the share of cost by service setting, the patient relationship code (PRC) on the trigger/reaffirming claim line.

All stakeholders, including those who did not qualify to receive a Field Test Report, could review a series of mock reports that were representative of each measure and reporting type. Other public documentation posted during field testing included: measure specifications for each measure (comprising a Draft Cost Measure Methodology document and a Draft Measure Codes List file), a Measure Development Process document, a Frequently Asked Questions document,

[26]CMS, "Restructured BETOS Classification System https://data.cms.gov/provider-summary-by-type-of-service/provider-service-classifications/restructured-betos-classification-system

and a Measure Testing Form (including reliability and validity data).[27] During field testing, Acumen conducted education and outreach activities for interested parties, including multiple office hours sessions with specialty societies, a publicly posted field testing webinar recording, and Quality Payment Program Help Desk support.

Education and Outreach

Acumen directly conducted outreach via email to tens of thousands of interested parties using a contact list developed through previous public engagement efforts, as well as CMS and Quality Payment Program (QPP) listservs. Acumen also emailed clinicians who received the field test reports via CMS's GovDelivery.

Acumen and CMS hosted two office hours sessions in January 2023 to provide an overview of field testing to specialty societies, discuss what information their members would be particularly interested in, and answer any questions. Across both office hours sessions, there were attendees from targeted specialty societies who are likely to have members who could be attributed the measure.

Acumen worked closely with QPP Service Center to respond to stakeholder inquiries during field testing and continued to answer questions after the feedback period ended.

Acumen and CMS hosted the public 2023 MACRA Cost Measures Field Testing webinar in January 2023, where interested parties could learn more about field testing and the measures.[28] The webinar presentation outlined: (i) the cost measure field testing project (ii) the measure development and re-evaluation processes, and (iii) field testing activities. There was also an opportunity to ask questions during the Q&A portion of the webinar. The webinar recording, slides, and transcript were then made available for the public to review.

5.1.2.3 Feedback on Measure Performance and Implementation

Clinician Expert Workgroup Meetings

Feedback from the Workgroup members were recorded throughout the meeting. More formal feedback was gathered using polls, typically requesting for votes on certain specifications or appropriateness of the measure. These polls were conducted following each meeting and on an ad hoc basis, as needed.

Field Testing

In total, Acumen received 64 survey responses and 19 comment letters, including from specialty societies representing large numbers of potentially attributed clinicians.

Survey responses and comment letters were collected via an online survey, which contained general and detailed questions on the reports themselves, questions on the supplemental documentation, and questions on the measure specifications.

5.1.2.4 Feedback from Measured Entities

Field Testing

The Field Testing Feedback Summary Report presents feedback gathered during the field testing period, including cross-measure feedback and measure-specific feedback.[29] The measure-specific feedback was used as the basis for the post-field testing refinements that

[27]The measure specifications, mock reports, Measure Development Process document, Frequently Asked Questions document, and testing documents are posted on the Cost Measures Information Page: https://www.cms.gov/medicare/quality/value-based-programs/cost-measures.

[28]MACRA Wave 4 Cost Measures Field Testing Webinar materials are available on the Quality Payment Program Webinar Library: https://qpp.cms.gov/about/webinars.

[29]CMS, "2023 Field Testing Feedback Summary Report," Cost Measures Information Page, https://www.cms.gov/files/document/field-testing-feedback-summary-report-23-wave-5.pdf.

were made to the measures. Overarching feedback about data that would be helpful for clinicians to receive was recorded and shared with CMS for future consideration. See Section 5.1.2.6 for post-field testing refinements made to the ESRD measure.

5.1.2.5 Feedback from Other Users

Person and Family Engagement

Acumen incorporated thoughtful input from patients and caregivers throughout the ESRD measure development process. Before each Clinician Expert Workgroup meeting, Person and Family Partners (PFPs) would provide input through focus groups and interviews to help inform the Workgroup's discussion. Attending PFPs would then present the findings for the Workgroup members, which would help shape the recommendations they made for the measure specifications. Some examples of feedback the PFP the clinician and non-clinician specialties involved in their care, including nephrologists, primary care providers, endocrinologists, nurse practitioners, and social workers. PFPs also noted comorbid conditions (e.g., diabetes, sleep apnea, coronary artery disease, mental health conditions, and other kidney conditions), a lack of care coordination, and poor adherence to their medications and treatment regimens.

5.1.2.6 Consideration of Feedback

Field Testing

Careful consideration was given to all feedback gathered during field testing, and several updates were made to the measure based on the recommendations of field testing commenters and the Clinician Expert Workgroup comprised of subject matter and measure-development experts. Acumen conducted analyses into potential adjustments that could be made to the measures to improve their ability to assess the intended clinician population.

After field testing, Acumen compiled the feedback provided through the surveys and comment letters into a measure-specific report, which was then provided to the Clinician Expert Workgroup, along with the empirical analyses to inform their discussion and evaluation of any refinements needed to ensure that the measure is capturing what it was intended to capture.

The changes to the ESRD measure made after consideration of field-testing analyses and stakeholder feedback are:

- Including services for lipid management and peritonitis
- Adding a risk adjustor for crash starts

5.2 Usability

5.2.1 Improvement

The measure has not yet been implemented, and as such has not had influence over performance. Our testing suggests that there is a sufficiently large difference in measure scores among clinicians to meaningfully determine a difference in performance. The potential for this measure to distinguish between good and poor performance is promising in its ability to encourage improvement in cost efficient care.

Additionally, the face validity results suggest that the Clinician Expert Workgroup believes the measure assess care within the influence of the clinician and can positively impact care provision and coordination.

5.2.2 Unexpected Findings

There were no unexpected findings during the development and testing of this measure. The measure has not been implemented at this time, so we do not have data that confirms

unexpected findings related to its implementation. However, Acumen considered the potential unintended consequences of having a cost measure for this clinical area (e.g., potential stinting in care to receive a better cost score). For instance, the empiric validity data previously presented in section 3.3 demonstrates that cost measure reflects the cost directly related to treatment choices and the cost of related adverse outcomes. For example, minor procedures and Part D medications are associated with adverse events, which suggest a potential overuse or frequent co-occurrence with adverse events.

Additionally, CMS monitors measures that are in use and has multiple processes in place to allow for changes to a measure if appropriate. These include i) annual maintenance for non-substantial changes and upkeep, ii) ad hoc maintenance if a specific issue occurs or a large change in clinical guidance takes place, and iii) measure reevaluation every three years where the suitability of a measure's specifications is comprehensively reassessed. If in the event the measure did have any unexpected findings, it would be identified and resolved through one of these methods.

5.2.3 Unexpected Benefits

Since the measure has not been implemented at this time, there are no testing results that identify unexpected benefits. However, many clinicians can only be assessed by the MSPB Clinician and TPCC measures in the cost performance category currently. This measure would provide a more tailored assessment of the care they have influence over, which many clinicians may prefer to be measured by compared to the population-based cost measures like MSPB Clinician or TPCC.

6.0 Related and Competing Measures

6.1 Relation to Other Measures

There are no competing measures with this measure. However, the following measures have been identified as potentially related.

Table 14. Quality Measures Potentially Relevant for the ESRD Episode Group

Measure Title	Measure ID	Measure Description	Measure Type
Controlling High Blood Pressure (CBP)	CMIT 167 (MIPS 236 & MVP M0002)	Percentage of patients 18-85 years of age who had a diagnosis of essential hypertension starting before and continuing into, or starting during the first six months of the measurement period, and whose most recent blood pressure was adequately controlled (<140/90mmHg) during the measurement period	Intermediate Outcome
Hemodialysis Vascular Access: Long-term Catheter Rate	CMIT 313 (MIPS 482, MVP M0002, ESRD QIP)	Percentage of adult hemodialysis patient-months using a catheter continuously for three months or longer for vascular access attributable to an individual practitioner or group practice.	Intermediate Outcome
Kidney Health Evaluation	CMIT 989 (MIPS 488)	Percentage of patients aged 18-75 years with a diagnosis of diabetes who received a kidney health evaluation defined by an Estimated Glomerular Filtration Rate (eGFR) AND Urine Albumin-Creatinine Ratio (uACR) within the measurement period.	Process

The quality measures listed above are related to the ESRD measure as they include metrics focused on similar patient cohorts, clinically related to the care provided for the episode group, or complementary care. While two quality measures are specific to kidney care, the remaining measure applies to a broader cohort of patients with high blood pressure.

6.2 Harmonization

During the measure's development, the Clinician Expert Workgroup specifically considered how to align relevant cost and quality measures (e.g., episode window length). This measure's development is aligned with episode-based cost measures currently used in the program. The ESRD measure was also developed in consideration of alignment opportunities with CMS' KCF and CKCC payment Options of the KCC Advanced Payment Model.

6.3 Competing Measures

There are no measures that conceptually address both the same measure focus and the same target population as the ESRD measure.

Additional Information

End-Stage Renal Disease Clinician Expert Workgroup Members:
As noted above, the following members provided detailed feedback on the measure specifications throughout its development based on public comments, clinical expertise, and empirical analyses.

- Donnie Batie, MD, FAAFP
- Peter Bustamante, MD
- Daniel Duzan, MD, SFHM, CPC
- Connie Hemeyer, MSN, APRN, FNP-BC
- Stephen Hohmann, MD, FACS
- Muralidharan Jagadeesan, MBBS, FACP, FASN
- Namirah Jamshed, MD
- Stephanie Jernigan, MD
- Daniel Lam, MD
- Alexander Liang, MD
- Devika Nair, MD, MSCI
- Connie Rhee, MD, MSc
- Jane Schell, MD
- Jeffrey Silberzweig, MD, FACP, FASN
- Joseph Vassalotti, MD
- Daniel Weiner, MD, MS

Measure Developer Updates and Ongoing Maintenance
The measure is not currently in use, but the earliest possible release of the measure in MIPS would be CY2025. If the measure becomes finalized for use in MIPS, it would undergo annual maintenance and a comprehensive re-evaluation every 3 years. This measure is included on the 2023 Measures Under Consideration (MUC) List and will be reviewed by PRMR in winter of 2023-2024. There are no further updates or reviews for this measure scheduled at this time.