Quality Payment
PROGRAM

Kidney Transplant Management

Measure Justification Form

December 2023

Disclaimer: The data reported here have been supplied by UNOS as the contractor for the Organ Procurement and Transplantation Network (OPTN). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government.

The study was submitted to a functioning IRB for review and approval. The IRB determined the need for an expedited review. Through this review, the IRB found no concerns with the research plan and provided approval to conduct the research project.

This study used data from the Organ Procurement and Transplantation Network (OPTN). The OPTN data system includes data on all donor, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN contractor.

Table of Contents

1.0 Introduction

This Measure Justification Form (MJF) provides results for the testing and evaluation of the Kidney Transplant Management 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 as of 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

Kidney Transplant Management Episode-Based Cost Measure

1.5 Type of Measure

Cost/Resource Use

1.6 Measure Description

The Kidney Transplant Management 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 related to kidney transplant, beginning 90 days post-transplant. This chronic condition measure includes the costs of services that are clinically related to the attributed clinician's role in managing care during a Kidney Transplant Management episode.

[1]CMS, "Kidney Transplant Management Measure Methodology," *Cost Measure Information Page,* https://www.cms.gov/medicare/quality/value-based-programs/cost-measures
[2]CMS, "Kidney Transplant Management Measure Codes List" *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 Kidney Transplant Management 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 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 measures and 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 costs of care 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.

As of 2020, an estimated 70% of the 800,000 patients with prevalent ESRD (including both new and preexisting ESRD) receive some type of maintenance dialysis, while around 30% are living with a functional kidney transplant.[4] Patients with ESRD who receive a kidney transplant tend to experience lower mortality over time and have an improved health-related quality of life.[5,6] At present, there are approximately 89,000 patients with ESRD on the Organ Procurement and Transplantation Network (OPTN) kidney donation waitlist.[7] The median ESRD patient spends more than four years on the waitlist, and about 24,000 patients undergoing dialysis received a kidney transplant in 2019.[4]

From 2010 to 2020, the total expenditures for beneficiaries with a kidney transplant increased by about 20% from $3.5 billion to $4.2 billion, and costs per person per year for a patient with transplant is about 40% that of patients undergoing hemodialysis.[4] Costs from post-transplant readmissions, medications, and delayed graft function or graft failure contribute significantly to overall healthcare costs for patients with a kidney transplant.

---

[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] 2022 United States Renal Data System (USRDS) Annual Data Report: Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases. 2022. https://usrds-adr.niddk.nih.gov/2022

[5] Tonelli M, Wiebe N, Knoll G, et al. Systematic review: kidney transplantation compared with dialysis in clinically relevant outcomes. *Am J Transplant*. 2011;11(10):2093-2109. doi:10.1111/j.1600-6143.2011.03686.x.

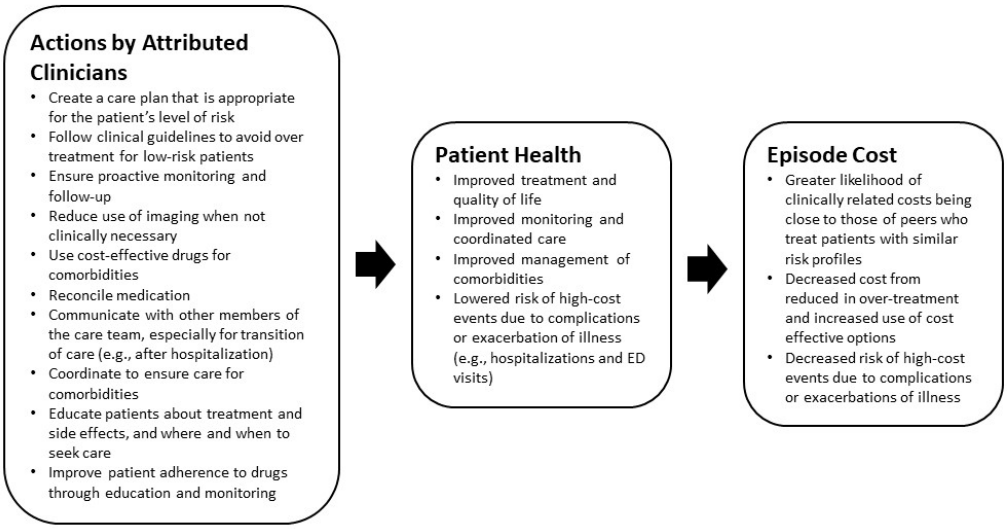
[6] Wolfe RA, Ashby VB, Milford EL, et al. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *N Engl J Med*. 1999;341(23):1725-1730. doi:10.1056/NEJM199912023412303.

[7] Organ Procurement and Transplantation Network (OPTN). National Data. Accessed April 20, 2023. https://optn.transplant.hrsa.gov/data/

The Kidney Transplant Management episode-based cost measure was initially recommended for development by the Clinician Expert Workgroup supporting development of the CKD and ESRD episode-based cost measures because assessing kidney transplant management was essential to capturing the entire care continuum for these patients with kidney disease. It was further supported by the public for these same reasons and because it represented a high-cost area that where there are strong opportunities of improvement for patient care. 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.

2.1.1 Logic Model

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



2.2 Performance Gap

2.2.1 Rationale

According to the literature and feedback received through stakeholder input activities, the Kidney Transplant Management measure's focus represents an area with opportunities for improvement. As discussed in the rest of this section, primary opportunities for improving cost outcomes associated with kidney transplant management include graft monitoring, reducing adverse patient outcomes, and reducing downstream costs. The literature scan identifies four critical areas for improving the costs related to kidney transplants:

1. Readmissions and emergency department (ED) visits
2. Immunosuppression regimens and medication adherence to prevent kidney rejection
3. Management of comorbidities including cardiovascular disease (CVD) management
4. Infection prevention and control

Post-transplant readmissions make up 20% of all Medicare payments for transplantation, but up to half of these readmissions are preventable.[8] Readmissions within 30 days that have the

[8] Hogan J, Arenson MD, Adhikary SM, et al. Assessing Predictors of Early and Late Hospital Readmission After Kidney Transplantation. *Transplant Direct*. 2019;5(8):e479. Published 2019 Jul 29. doi:10.1097/TXD.0000000000000918.

highest costs are those associated with post-operative complications, transplant rejection, and infection,[9] and multiple studies show that the majority of kidney transplant recipients will visit ED in the first two years post transplantation.[10] Transplant recipients with longer initial hospital stays had 46% higher adjusted ED visit rate, suggesting that complications arise during that initial hospitalization may necessitate subsequent ED care.[11]

Management of comorbidities is costly to Medicare and accounts for a significant share of Medicare Part D drug spending for kidney transplant patients. Although recent studies reported stabilizing and/or decreasing trends of major adverse cardiovascular events, hospitalizations, and incidence of heart failure, cardiovascular disease remains a major contributor to health care utilization among kidney transplant recipients, and a major driver of graft loss.[12,13] A 2022 review showed that one in 10 kidney transplant recipients are diagnosed with post-transplant pulmonary hypertension within three years, which is associated with an increased risk of graft failure.[12] Other risk factors associated with poor graft outcomes include symptomatic atherosclerotic vascular disease and heart failure.[14]

Similarly, infections also pose a significant burden to patients and impact healthcare costs. Cytomegalovirus (CMV) is one of the most common viral opportunistic infections, and immunosuppressive medication increases the risk for infection with CMV.[15] Ureteric stents, which are routinely used in kidney transplantation to reduce major urological complications are also associated with a rise in urinary tract infections (UTIs). A review of 141,661 kidney transplant recipients from over a 15-year period found the cumulative incidence of post-transplant infection was 36.9% at three months, 53.7% at one year, and 78.0% at five years, with nearly half (46.8%) of the infections occurring in the urinary tract.[16] Some studies have indicated the ureteric stent benefit from the reduction in urological complications is outweighed by the risk of UTIs while immunocompromised.[17] Cost increase in the first year post-transplant for UTI alone has been estimated at $17,691.[18]

---

[9] Famure O, Kim ED, Au M, et al. What Are the Burden, Causes, and Costs of Early Hospital Readmissions After Kidney Transplantation? *Prog Transplant*. 2021;31(2):160-167. doi:10.1177/15269248211003563.
[10] Schold JD, Elfadawy N, Buccini LD, et al. Emergency Department Visits after Kidney Transplantation. *Clin J Am Soc Nephrol*. 2016;11(4):674-683. doi:10.2215/CJN.07950715.
[11] Dalrymple LS, Romano PS. Emergency Department Visits after Kidney Transplantation. *Clin J Am Soc Nephrol*. 2016;11(4):555-557. doi:10.2215/CJN.02040216.
[12] Goyal A, Chatterjee K, Mathew RO, et al. In-Hospital Mortality and Major Adverse Cardiovascular Events after Kidney Transplantation in the United States. *Cardiorenal Med*. 2019;9(1):51-60. doi:10.1159/000492731.
[13] Lenihan CR, Liu S, Airy M, Walther C, Montez-Rath ME, Winkelmayer WC. The Association of Pre-Kidney Transplant Dialysis Modality with de novo Posttransplant Heart Failure. *Cardiorenal Med*. 2021;11(5-6):209-217. doi:10.1159/000518535.
[14] Otunla AA, Shanmugarajah K, Salciccioli JD, et al. Symptomatic atherosclerotic vascular disease and graft survival in primary kidney transplant recipients - Observational analysis of the united network of organ sharing database. *Transpl Immunol*. 2022;75:101734. doi:10.1016/j.trim.2022.101734.
[15] Jun KW, Lim JH, Hwang JK, et al. Selection of More Vulnerable Patients for Cytomegalovirus Infection in Renal Transplant Recipients with Antithymocyte Globulin Induction Therapy: An Analysis of Risk Factors and Cell-Mediated Immunity. *Transplant Proc*. 2021;53(7):2252-2260. doi:10.1016/j.transproceed.2021.07.031.
[16] Jackson KR, Motter JD, Bae S, et al. Characterizing the landscape and impact of infections following kidney transplantation. *Am J Transplant*. 2021;21(1):198-207. doi:10.1111/ajt.16106.
[17] Mosqueda AO, Hernández EEL, Morales GC, et al. Association Between the Placement of a Double-J Catheter and the Risk of Urinary Tract Infection in Renal Transplantation Recipients: A Retrospective Cohort Study of 1038 Patients. *Transplant Proc*. 2021;53(6):1927-1932. doi:10.1016/j.transproceed.2021.05.002.
[18] Naik AS, Dharnidharka VR, Schnitzler MA, et al. Clinical and economic consequences of first-year urinary tract infections, sepsis, and pneumonia in contemporary kidney transplantation practice. *Transpl Int*. 2016;29(2):241-252. doi:10.1111/tri.12711.

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).

The score interquartile range (IQR) for both TINs and TIN-NPIs is greater than 25 percent of the mean score. Additionally, for both TINs and TIN-NPIs, the 90th percentile score was much greater than the 10th percentile score; the 90th percentile score was more than 60% higher than the 10th percentile at the TIN-level, and almost 90 percent higher at the TIN-NPI level. The distributions show meaningful variation in cost performance and suggest that there's room for improvement in the costs of care for a kidney transplant management episode.

Table 1. Distribution of the Measure Score

Metric	TIN	TIN-NPI
Count	696	525
Mean Score	$20,827	$20,699
Score Standard Deviation	$4,186	$5,076
Minimum Score	$9,796	$9,983
Maximum Score	$38,247	$39,834
Score Interquartile Range (IQR)	$5,353	$6,596
Score Percentile		
10th	$15,916	$14,818
20th	$17,425	$16,317
30th	$18,556	$17,571
40th	$19,689	$18,893
50th	$20,593	$20,106
60th	$21,467	$21,153
70th	$22,681	$22,677
80th	$23,953	$24,806
90th	$26,001	$27,989

2.2.3 Disparities

Sections 3.1.7 and 3.5.5 provide data on how the measure, as specified, addresses disparities in care related to kidney transplant management.

3.0 Scientific Acceptability

3.1 Data Sample Description

Testing is based on the full population of measured entities meeting the testing volume threshold (20 episodes) and inclusion and exclusion criteria for the measure, unless otherwise stated. The testing is not based on a sample of the population.

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), Medicare Enrollment Database (EDB), and additional registry data for kidney transplant characteristics.

3.1.2 Specific Dataset Used for Testing

The Kidney Transplant Management measure uses Medicare Parts A, B, and D claims data maintained by CMS. Part A, B, and D claims data are used to build episodes of care, calculate episode costs, and construct risk adjustors. 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 health care 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 MDS. Specifically, the MDS is used to create the long-term care indicator variable in risk adjustment. Registry data are used to inform risk adjustment variables related to kidney transplant characteristics.

3.1.3 Dates of the Data Used in Testing

Kidney Transplant Management 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 Kidney Transplant Management measure.

Table 2: Measured Entities Demographics

Metric	TIN		TIN-NPI	
No data	Count	%	Count	%
Count	696	100%	525	100%
Number of Episodes Attributed	-	-	-	-
20-39 Episodes	336	48.28%	341	64.95%
40-59 Episodes	105	15.09%	130	24.76%
60-79 Episodes	61	8.76%	36	6.86%
80-99 Episodes	41	5.89%	9	1.71%
100-199 Episodes	77	11.06%	8	1.52%

Metric	TIN		TIN-NPI	
No data	Count	%	Count	%
200-299 Episodes	36	5.17%	1	0.19%
300+ Episodes	40	5.75%	0	0%
Census Region	-	-	-	-
Northeast	128	18.39%	121	23.05%
Midwest	152	21.84%	89	16.95%
South	257	36.93%	196	37.33%
West	156	22.41%	117	22.29%
Unknown	3	0.43%	2	0.38%

3.1.6 Patient Cohort Included in the Testing and Analysis

Table 3 shows the patient population for the Kidney Transplant Management measure testing. It consists of Medicare beneficiaries enrolled in Medicare Parts A and B who received care related to kidney transplant management that triggers a Kidney Transplant Management episode.

Table 3: Beneficiary Demographics

Metric	Value
Count	53,945
Mean Age	62.08 years
Female %	41.76%
Part D Enrollment %	79.41%

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.[19]

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[19]	• 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[19]	No

[19] 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[20,21]	
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[22]	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 Kidney Transplant Management 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. CMS continues to perform corrective actions and give providers additional education to ensure accurate billing.

[20] 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).
[21] 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).
[22] 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

- 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.

Entity-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 used a signal-to-noise analysis.

This approach seeks to determine how much of the variation in the measure score is explained by differences among clinicians' 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 one 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.[23] The fiscal year 2022 Medicare fee-for-service program proper payment rate was 92.5%.[24]

Entity-level Reliability

The table below shows reliability metrics at the 20-episode testing volume thresholds. While higher thresholds generally yield higher reliability results, these increases must be considered against decreasing the number of clinicians and clinician groups eligible for the measure, which would limit the applicability of measures to larger group practices and potentially limit the impact of the measure in encouraging performance improvement. 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 with consideration of previously established policies for entity-level reliability for MIPS cost measures.

[23]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.
[24]Ibid.

Table 5: Reliability at the Entity Level

Reporting Level	Entities Meeting Testing Minimum	Mean Reliability	Median Reliability	% Above 0.4	% Above 0.7
TIN	696	0.43	0.39	48.99%	12.36%
TIN-NPI	525	0.42	0.41	51.05%	4.38%

3.2.4 Interpretation

The results of the data element testing show very high reliability of the critical data elements used by the measure. At the entity level, the measure is moderately reliable for both the TIN and TIN-NPI reporting levels, at 0.425 and 0.418, respectively. 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.

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 Kidney Transplant Management 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 Kidney Transplant Management Clinician Expert Workgroup;
(ii) a Technical Expert Panel (TEP);
(iii) the Person and Family Partners; and
(iv) other interested parties through the National Field Testing public comment period.

This process is detailed in the Episode-Based Cost Measures Development Process document posted on the Cost Measure Information Page.[25]

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 evaluated on services and costs that are clinically related to the attributed clinician's role in managing care during a Kidney Transplant Management episode, thus limiting cost variation unrelated to clinician care in this measure. Therefore, assigned services are services that the

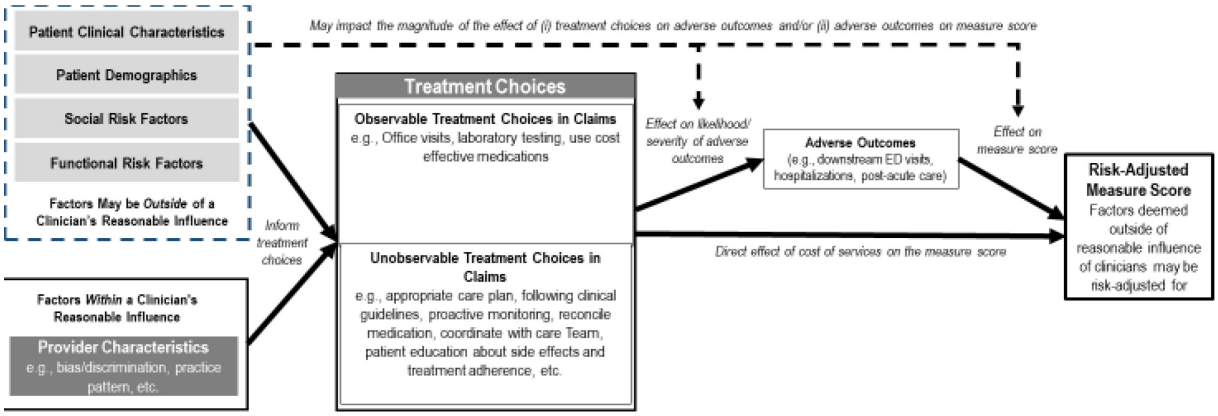
[25] CMS, Cost Measure Information Page, https://www.cms.gov/medicare/quality/value-based-programs/cost-measures.

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 Kidney Transplant Management 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 (O/E) 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	0.02 [0.02,0.02] (p < 0.01)	-	0.02 [0.02,0.03] (p < 0.01)	-
Dialysis	0.05 [-0.10,0.19] (p = 0.53)	1.37 [-1.72,4.45] (p = 0.39)	-0.05 [-0.21,0.10] (p = 0.50)	-1.77 [-4.79,1.25] (p = 0.25)
Outpatient Evaluation & Management Services	-0.03 [-0.07,0.01] (p = 0.13)	3.68 [2.88,4.48] (p < 0.01)	-0.06 [-0.10,-0.01] (p = 0.01)	2.73 [1.90,3.56] (p < 0.01)
Major Procedures	0.10 [0.02,0.19] (p = 0.02)	-3.46 [-5.30,-1.62] (p < 0.01)	0.07 [-0.04,0.18] (p = 0.22)	-2.42 [-4.55,-0.28] (p = 0.03)
Ambulatory/Minor Procedures	0.03 [0.01,0.06] (p < 0.01)	1.18 [0.66,1.70] (p < 0.01)	0.02 [-0.01,0.05] (p = 0.11)	0.50 [-0.06,1.06] (p = 0.08)
Anesthesia Services	0.13 [-0.23,0.49] (p = 0.47)	26.27 [18.90,33.65] (p < 0.01)	0.11 [-0.34,0.56] (p = 0.63)	31.20 [22.94,39.47] (p < 0.01)
Laboratory, Pathology, and Other Tests	0.04 [0.02,0.06] (p < 0.01)	-0.23 [-0.57,0.10] (p = 0.17)	0.03 [0.01,0.04] (p < 0.01)	-0.14 [-0.39,0.12] (p = 0.30)
Imaging Services	0.12 [0.03,0.22] (p = 0.01)	2.59 [0.58,4.61] (p = 0.01)	0.14 [0.02,0.26] (p = 0.02)	1.40 [-0.87,3.66] (p = 0.23)
Durable Medical Equipment and Supplies	0.04 [-0.03,0.11] (p = 0.24)	2.53 [1.10,3.96] (p < 0.01)	-0.04 [-0.21,0.13] (p = 0.63)	0.68 [-2.59,3.94] (p = 0.69)
Chemotherapy and Other Part B-Covered Drugs	0.02 [0.02,0.03] (p < 0.01)	-0.07 [-0.15,0.02] (p = 0.13)	0.04 [0.04,0.04] (p < 0.01)	0.00 [-0.08,0.07] (p = 0.93)
Part-D Drugs	0.01 [0.01,0.02] (p < 0.01)	0.23 [0.04,0.42] (p = 0.02)	0.01 [0.00,0.01] (p = 0.19)	0.09 [-0.08,0.26] (p = 0.29)

3.3.4 Interpretation

Overall, the results demonstrate that the cost measure is reflective of both the cost directly related to treatment choices, as well as cost of adverse outcomes as a result of care (Table 6). Therefore, there's evidence that the measure is capturing what it purports to measure.

Model 1 shows that the cost of adverse events (e.g., hospitalizations, emergency department visits, or post-acute care that are clinically related to kidney transplant management) is associated with a slightly worse measure score for both TIN and TIN-NPI. While dialysis cost

may be high in nominal amount, the results suggest that it is not a significant cost driver of the measure score. Outpatient evaluation and management services are associated with a better score, but the result is only statistically significant at the TIN-NPI level.

Major or minor procedures, anesthesia, imaging, durable medical equipment, and Parts B or D medications are associated with a worse measure score. However, major procedures are associated with lower costs of adverse events, which suggest that avoiding these services may not be good candidates for improving overall cost performance. Additionally, minor procedures and anesthesia are associated with high cost of adverse events, which can either indicate frequent cooccurrence with adverse events or a potential of overuse. While the results are less clear at the TIN-NPI level, imaging services and durable medical equipment are associated with both lower measure scores and higher cost of adverse events, which suggest it may be an area for cost improvement. Parts B or D medications do not show a statistically signification association with adverse events, which suggest that their association with a lower measure score may be due to a potential of overuse or inefficient selection.

3.4 Exclusions Analysis

3.4.1 Method of Testing Exclusions

Exclusions are used in the Kidney Transplant Management measure to ensure a comparable patient population within the scope of the measure's focus on care related to kidney transplant 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.
- Measure-specific exclusions including atypical hemolytic-uremic syndrome (HUS)
    - This episode is excluded because these episodes can be clinically distinct from the overall kidney transplant management 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 Kidney Transplant Management measure is provided in the Measure Codes List available on the Cost Measure Information Page.[26]

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

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

exclusion criteria ensure that the reportable episode populations are more homogenous and comparable than all episodes meeting triggering logic.

Table 7: Cost Statistics for Measure Exclusions

Exclusion	Episodes		Observed Cost					
			Mean	Percentile				
	#	% of All Episodes Meeting Triggering Logic		10th	25th	50th	75th	90th
All Episodes Meeting Triggering Logic	96,940	100%	$29,871	$3,319	$6,413	$14,906	$33,827	$68,749
Episode Length Less Than 1 year	4,957	5.11%	$116,333	$9,672	$25,489	$65,841	$143,228	$263,502
Beneficiary Death in Episode	8,451	8.72%	$93,751	$13,002	$29,129	$61,383	$114,404	$196,265
Outlier	1,720	1.77%	$69,779	$10,078	$20,625	$92,464	$116,285	$116,285
No Attributed TIN-NPI	18,107	18.68%	$36,368	$3,891	$7,921	$18,594	$40,781	$80,219
TIN does not Meet Testing Volume Threshold	27,858	28.74%	$31,720	$3,515	$6,736	$15,343	$35,385	$74,981
TIN-NPI does not Meet Testing Volume Threshold	57,752	59.57%	$29,083	$3,242	$6,178	$14,273	$33,003	$68,453
Atypical Hemolytic Uremic Syndrome (HUS)	217	0.22%	$309,123	$11,010	$32,845	$300,901	$502,114	$702,485
Reportable Episodes (if all clinicians reported as TIN at the Testing Volume Threshold)	60,682	62.60%	$20,544	$3,036	$5,699	$12,716	$27,349	$48,890
Reportable Episodes (if all clinicians reported as TIN-NPI at the Testing Volume Threshold)	19,202	19.81%	$19,762	$2,996	$5,649	$12,378	$26,231	$46,181

3.4.3 Interpretation

The statistical results show that the reportable episode populations are more homogenous and comparable than all episodes meeting triggering logic, supporting the exclusion of these episodes to ensure a comparable patient cohort that will yield a clinically coherent measure and meaningful information to attributed clinicians. Further discussion of the results for exclusions applied based on the clinical validity of the study population are provided below.

Overall, exclusion criteria decrease the distribution of observed costs from the mean of $29,871 of all episodes meeting trigger logic to $20,544 for reportable episodes at the TIN-level and $19,762 for reportable episodes at the TIN-NPI level. The excluded cohorts have higher mean observed costs compared to all episodes meeting triggering logic. The largest exclusions come from applying the testing volume threshold to ensure a sufficient sample size for the measure.

Episodes shorter than one year are excluded because the methodology for the chronic measures requires at least one year of claims data to measure clinician cost performance to

ensure sufficient observation of chronic care, which is often intermittent and sparse over a long period of time. Although these episodes are excluded during the performance period being examined, they are likely to be included in the following performance period once the episode length is longer than one year. Episodes where a beneficiary died before the episode end date are excluded because they do not provide sufficient data in the episode window period. These episodes also have a higher mean observed cost than all episodes meeting triggering logic, at $93,751. 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 $69,779 compared to $29,871 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 $10,078 and at the 90th percentile the observed cost is $116,285.

Based on the input from the clinical expert workgroup, atypical hemolytic uremic syndrome (HUS) is excluded because these episodes can be clinically distinct from the overall kidney transplant management population. Episodes with atypical HUS have mean observed costs that are at least 12 times higher than all episodes meeting trigger logic (Table 14), which suggests that they may have distinct resource use patterns from typical kidney transplant management.

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 139 risk factors and stratification by 2 risk categories.

The risk adjustment model for the Kidney Transplant Management 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 other risk factors.

The model also includes measure-specific factors:

- Characteristics of the transplanted kidney (Kidney Donor Profile Index [KDPI] for transplanted kidneys from deceased donors, transplanted kidney from blood type incompatible donor, Human Leukocyte Antigen (HLA) donor/recipient mismatch, age of the transplant)
- Highly sensitized patients
- An episode that ended in renal failure
- Other patient health status factors prior to episode (heart failure hospitalization, glomerulonephritis, lupus nephritis, BK nephropathy, cytomegalovirus (CMV) viremia, other organ transplant, prior transplant rejection treatment, two or more prior kidney transplants).

A separate linear regression is run for each sub-group, i.e., episodes with and without 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 Cost Measure Information Page.[27]

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 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 Kidney Transplant Management 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 3 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.[19,28,29,30,31] 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

[27]CMS, Cost Measure Information Page, https://www.cms.gov/medicare/quality/value-based-programs/cost-measures.
[28]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.
[29]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
[30]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/.
[31] 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

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 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[32] and the Report to Congress: Risk Adjustment in Medicare Advantage[33]. 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.

[32]Pope, Gregory C., John Kautter, et al., "Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report." RTI International: March 2011.
[33]CMS, "Report to Congress: Risk Adjustment in Medicare Advantage," https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/RTC-Dec2018.pdf.

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

Subgroup Risk Model	% 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 with Part D Enrollment	79.58%	0.16 (p: <0.001)	0.15 (p: <0.001)	0.16 (p: <0.001)
TIN without Part D Enrollment	20.42%	0.11 (p: 0.26)	0.12 (p: 0.22)	0.04 (p: 0.73)
TIN-NPI with Part D Enrollment	79.84%	0.17 (p: <0.001)	0.14 (p: <0.001)	0.15 (p: <0.001)
TIN-NPI without Part D Enrollment	20.16%	0.18 (p: 0.13)	0.2 (p: 0.09)	0.15 (p: 0.36)

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 Episode	Dual Episodes	Non-Dual Episodes	All Episodes	Dual Episodes	Non-Dual Episodes
ALL	1.00	1.03	0.99	1.02	1.05	1.01
0%-20%	1.00	0.99	1.00	0.97	1.00	0.97
21%-40%	1.00	1.05	0.99	1.04	1.06	1.03
41%-60%	1.00	1.04	0.99	1.03	1.08	1.01
61%-80%	1.00	1.03	0.98	1.02	1.05	1.00
81%-100%	1.02	1.05	0.98	1.03	1.04	1.03

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	6%	93%	1%
TIN-NPI	6%	93%	2%

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	77.4%	13.5%
TIN-NPI	81.1%	9.5%

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). While many clinicians are able to perform equally well on their dual episodes and non-dual episodes, there are many more clinicians who are performing significantly worse on their dual episodes than their non-dual episodes, which suggests that clinicians aren't able to fully mitigate the effect of SRFs (Table 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 Kidney Transplant Management cost measure, calculated by dividing explained sum of squares by total sum of squares is 0.269. The adjusted R-squared is 0.265. More information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.[34]

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 minimal variation among risk deciles, as predictive ratios range from 0.90 and 1.06 across all risk deciles (with an overall average of 1.02).

Table 12: Predictive Ratio by Decile of Predicted Episode Cost

Decile	Average Predictive Ratio
Decile 1	0.90
Decile 2	0.97
Decile 3	0.93
Decile 4	0.96

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

Decile	Average Predictive Ratio
Decile 5	0.98
Decile 6	0.97
Decile 7	1.01
Decile 8	1.02
Decile 9	0.99
Decile 10	1.06

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.[35] 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 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 demonstrates that the risk adjustment model is consistent, with the average predictive ratios observed to be close to 1.00 across all deciles, with the range between 0.90 and 1.06. Overall, the risk adjustment model does not over- or under-predict cost across the full range of resource use patterns in the population.

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 (Table 1). Additionally, the testing results found that the rates of inpatient admission and emergency room visit during an episode are 31.7% and 42%, respectively.

3.6.3 Interpretation

There is substantial variation observed in the measure score in both TIN and TIN-NPI levels, indicated by the interquartile ranges, standard deviations, and interquartile range. The magnitude of the observed variation is in the thousands of dollars, which indicates that there are

[35]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.

opportunities to close the gaps between the most and least efficient clinicians. The rates of inpatient admission and emergency room visit during an episode suggest a substantial opportunity for improvement, which added over $20,000 and almost $9,000 per episode, respectively, compared to an average episode.

3.7 Missing Data Analysis and Minimizing Bias

3.7.1 Method

Since CMS uses Medicare claims data to calculate the Kidney Transplant Management 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 Kidney Transplant Management 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 Kidney Transplant Management 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 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
No Continuous Enrollment in Medicare Parts A and B, and Any Enrollment in Part C	26,689	$14,308	$491	$1,600	$5,353	$15,021	$32,256
Beneficiary Resides Outside of U.S. or its Territories	268	$16,552	$987	$1,898	$6,634	$17,290	$37,136
Primary Payer Other than Medicare	25,292	$21,018	$1,009	$2,733	$8,430	$22,412	$48,809

3.7.3 Interpretation

The table above (Table 13) presents three distinct missing data categories, each with its associated average costs and variability. In the first category, where individuals lack continuous enrollment in Medicare Parts A and B but are enrolled in Part C, the average cost is $14,308. The costs within this group vary widely, from $491 at the 10th percentile to $32,256 at the 90th percentile. The second category, involving Medicare beneficiaries who live outside the U.S. or its territories, has an average cost of $16,552. Within this category, costs range from $987 at

the 10th percentile to $37,136 at the 90th percentile. The third category comprises individuals with a primary payer other than Medicare. This group has a mean cost of $21,018, with costs ranging from $1,009 at the 10th percentile to $48,809 at the 90th percentile. Therefore, it is appropriate to remove these episodes as they are likely indicators of a discontinuation of the patient-clinician relationship or an absence of Medicare usage, and therefore do not provide sufficient data during the episode window. Given their limited frequencies, the impact of removing these episodes on the overall measure should be minimal while ensuring that clinicians are fairly evaluated on episodes with complete information.

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-rulemaking 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 Kidney Transplant Management 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 Kidney Transplant Management 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.[36] 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 1,239 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.

[36]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)[37]
    - 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,

[37]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).[38] 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.[39] 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 48 survey responses and 5 comment letters, including from specialty societies representing large numbers of potentially attributed clinicians and from persons with lived experiences.

Survey responses and comment letters were collected via two online surveys, 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.[40] The

[38]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.

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

[40]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.

measure-specific feedback was used as the basis for the post-field testing refinements that 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 Kidney Transplant Management measure.

5.1.2.5 Feedback from Other Users

Person and Family Engagement

Acumen incorporated thoughtful input from patients and caregivers throughout the Kidney Transplant Management 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 include the types of services that they typically received and what helped to improve their care (e.g., lab testing, medication management, durable medical equipment) and noted the types of clinicians that contributed to their kidney care team (e.g., nephrologists, family internists, transplant coordinators). They also highlighted areas of concerns, such as complications and lack of care coordination that impacted the quality of their care.

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 Kidney Transplant Management measure made after consideration of field-testing analyses and stakeholder feedback are:

- Adjusted for factors affecting kidney quality
- Conducted additional investigations on potential impact of risk adjustment for prior transplant rejection treatment

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 did consider 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 example, the empiric validity data previously presented in section 3.3 demonstrates that while providing the cost of dialysis may be very high, it is not a major driver of the measure score and, therefore, underscoring the robustness of the measure in differentiating performance that is most relevant to chronic management patients who received kidney transplant.

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 Kidney Transplant Management Episode Group

Measure Title	Measure ID	Measure Description	Measure Type
Kidney Health Evaluation Episode-Based Cost Measure	989	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
Adult Kidney Disease: Angiotensin Converting Enzyme (ACE) or Angiotensin Receptor Blocker (ACE) Therapy	777	Percentage of patients aged 18 years and older with a diagnosis of CKD (Stages 1-5, not receiving Renal Replacement Therapy (RRT)) and proteinuria who were prescribed ACE inhibitor or ARB therapy within a 12-month period.	Process
Optimal ESRD Starts	NQF 2594	Optimal End Stage Renal Disease (ESRD) Starts is the percentage of new adult ESRD patients during the measurement period who experience a planned start of renal replacement therapy by receiving a preemptive kidney transplant, by initiating home dialysis, or by initiating outpatient in-center hemodialysis via arteriovenous fistula or arteriovenous graft.	Process

The MIPS quality measures listed above are related to the Kidney Transplant Management measure as they include metrics that focus on similar patient cohorts, are clinically related to the care provided for the episode group, or assess complementary care that may not be directly captured by the cost measure.

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 cost measure aligns with the Chronic Kidney Disease and End-Stage Renal Disease episode-based cost measures that were developed concurrently to ensure that there are cost measures to assess the overall care continuum for a patient with kidney disease and to ensure aligned incentives for care coordination.

6.3 Competing Measures

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

Additional Information

Kidney Transplant Management 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.

Kenneth Andreoni, MD, FACS, Director of Kidney Transplant
Sharon Bartosh, MD, Professor and Division Chief
Patricia Bartzak, DNP, RN, CMSRN, TCRN, CNRN, Staff Nurse
John Ducker, MD, Transplant Nephrologist
Muralidharan Jagadeesan, MBBs, FASP, FASN, Medical Director
Vineeta Kumar, MD, Professor
Krista Lentine, MD, PhD, Professor
Alexander Liang, MD, President/CEO
Sabrena McCarley, MBA-SL, OTR/L, CLIPP, RAC-CT, QCP, FAOTA, Director
Clifford Miles, MD, Transplant Nephrologist
Linda Moore, PhD, RDN, CCRP, Associate Research Professor and Associate Professor
Stephen Pastan, MD, Professor and Medical Director
Jeffrey Silberzweig, MD, FACP, FASN, Chief Medical Officer and Associate Professor
Daniel Weiner, MD, MS, Nephrologist

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.