Quality Payment
PROGRAM

Depression

Measure Justification Form

September 2022

Table of Contents

1.0 Introduction

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

1.1 Project Title

Physician Cost Measure and Patient Relationship Codes

1.2 Date

Information included is current on September 27, 2022.

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

1.4 Measure Name

Depression Episode-Based Cost Measure

1.5 Type of Measure

Cost/Resource Use

1.6 Measure Description

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

[1]CMS, "Depression Measure Methodology" and "Depression Measure Codes List," *MACRA Feedback Page,* https://www.cms.gov/Medicare/Quality-Payment-Program/Quality-Payment-Program/Give-Feedback

2.0 Importance

2.1 Evidence to Support the Measure Focus

The Depression 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 4 areas:

- quality
- improvement activities
- Promoting Interoperability
- cost

Each category assesses different aspects of care, and the categories are weighted such that they're combined into one composite score. CMS is introducing MIPS Value Pathways (MVPs) as a way 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 costs for patients.

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 of 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 healthcare expenditures' variation due to their ability to affect costs[2]. A cost measure offers 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.

The Depression episode-based cost measure was recommended for development through feedback gathered during a public comment period. The public recommended this measure because depression is a very prevalent condition compared to other mental illnesses, so its inclusion represents a larger gain for public health. A measure-specific Clinician Expert Workgroup was then convened with clinicians, healthcare experts, and patient representatives who have appropriate experience, to provide extensive, detailed input on this measure throughout its development

Depression affects 8.5% of Medicare beneficiaries.[3] While an estimated 5% of the Medicare population has Major Depressive Disorder (MDD), that rate increases to 5-10% in primary care

[2]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.

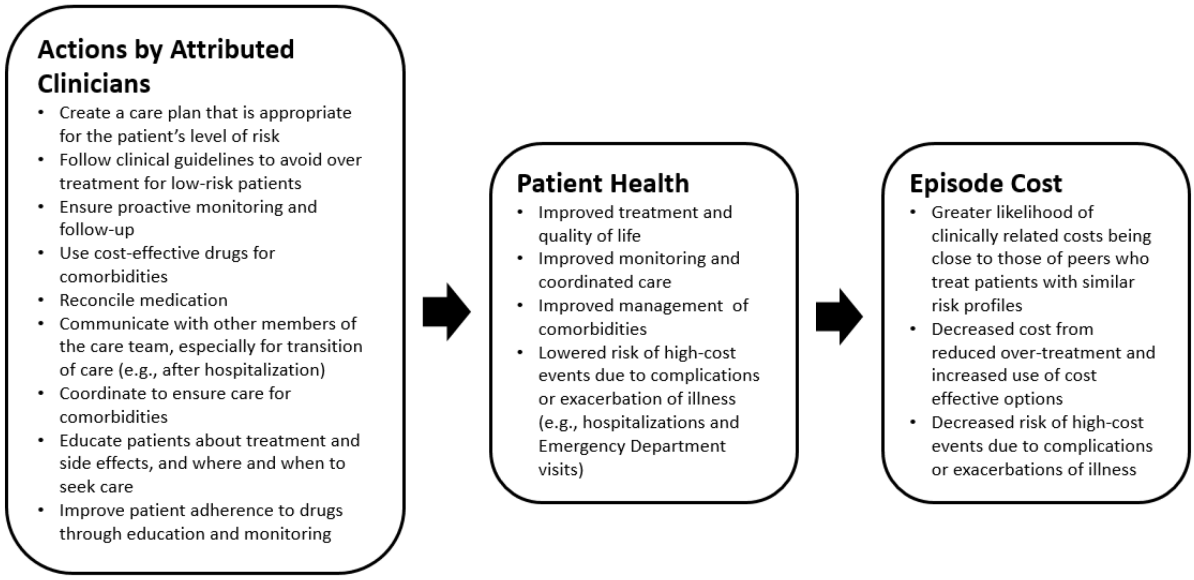
[3] Balasuriya L, Quinton JK, Canavan ME, et al. The Association Between History of Depression and Access to Care Among Medicare Beneficiaries During the COVID-19 Pandemic. Journal of General Internal Medicine. 2021; 36(12): 3778-3785

settings and 10-42% in inpatient settings.[4] The prevalence of MDD is higher among women compared to men, and highest among patients 90 years or older.[5]

One study estimated that the direct spending on mental health services for the Medicare population totaled $2.7 billion in 2015, or 4.2% of total spending; however, an extra 8.5% of additional medical spending was associated with mental illness, bringing the total spending associated with mental health disorders to 12.7%.[6] More specifically related to depression, one study indicated that the economic burden of MDD increased by 37.9% in the U.S. between 2010 and 2018, from $236.6 billion to $326.2 billion.[7]

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

This measure represents an opportunity to assess clinician performance related to managing chronic depression in the MIPS cost performance category, a clinical area where opportunities for improvement have been identified. It was developed with extensive input from clinical experts and other interested parties, including feedback from a public comment period, Technical Expert Panel (TEP), clinician expert workgroup, and patient/caregiver perspectives, as discussed above. The measure's development is aligned with episode-based cost measures currently used in the program.

[4] McQuaid, JR, Lin EH, Barber JP, et al. 2019. Clinical Practice Guideline for the Treatment of Depression Across Three Age Cohorts. American Psychological Association, Guideline Development Panel for the Treatment of Depressive Disorders.
[5] Bashyal R, Du H, Wang L, Yuce H, Baser O. PMH17 – Mortality and Prevalence of Major Depressive Disorder in the US Medicare Population from 2008-2013. Value in Health. 2016; 19(3): A184.
[6] Figueroa J, Phelan J, Orav J, et al. Association of Mental Health Disorders With Health Care Spending in the Medicare Population. JAMA Network Open; 2020; 3(3):e201210.
[7] Greenberg PE, Fournier AA, Sisitsky T, et al. The Economic Burden of Adults with Major Depressive Disorder in the United States (2010 and 2018). Pharmacoeconomics. 2021; 39(6):653-665.

According to the literature and feedback received through public comments and other input activities, this measure's focus represents an area where there are opportunities for improvement. As discussed in the rest of this section, the primary opportunities for improving Depression cost outcomes include (i) reducing the spending gap in depression care compared to other medical conditions, (ii) medication adherence, and (iii) integration of primary care and mental healthcare.

Existing literature reports that patients with depression are more likely to use healthcare services and resources for other types of medical illness beyond just mental health disorders compared to patients without depression.[8] One study found that the average total healthcare costs were higher in every component of care (i.e., primary care, emergency department visits, specialty medical visits, outpatient care) among patients with depression aged 60 or older compared to those without depression.[9] One study looking specifically at Medicare patients found that patients with a serious mental illness (which includes MDD) spent over one-third more on non-mental health conditions compared to those with no mental illness.[10] This may be due to the fact that individuals with depression or mental health disorders are more likely to have co-occurring chronic medical conditions that could be harder to manage, and thus, may result in more emergency department visits and hospitalizations.

Previous research also indicates that over half of patients with MDD don't adhere to prescribed medications (i.e., antidepressants), both in the primary care and psychiatric settings.[11] While non-adherence could be patient-related (i.e., due to concerns about side effects, cultural issues, costs[12,13]), there are also clinical considerations that play a factor, such as inadequate patient education, lack of shared decision-making, and lack of follow-up.[14] This points to the importance of and opportunity for clinicians to identify/recognize barriers to medication adherence and develop targeted interventions that address these barriers to increase patients' adherence. Effective management may include close monitoring of patients, proper and continued communication between clinicians and patients, involvement of caregivers or family members in treatment plans, and patient education to encourage adherence to medications and potential improved outcomes. These sentiments have also been echoed in Acumen's Person and Family Engagement (PFE) input processes, where patients and caregivers have emphasized the need for clear communication with patients, consideration of treatment goals, and thorough discharge planning after any admissions.

Several areas of research have indicated how the integration of primary care and mental healthcare can foster improvements in the management and treatment of patients with mental health disorders and chronic conditions as well as significantly reduce related spending. Complementing earlier cited research on how older patients with depression have

[8] Zivin K, Wharton T, Rostant O. The Economic, Public Health, and Caregiver Burden of Late-Life Depression. Psychiatric Clinics of North America. 2013; 36(4): 631-649.
[9] Katon WJ, Lin E, Russo J, et al. Increased Medical Costs of a Population-Based Sample of Depressed Elderly Patients. JAMA Psychiatry. 2003; 60(9):897-903.
[10] Figeuroa JF, Phelan J, Orav J, et al. Association of Mental health disorders with Health Care Spending in the Medicare Population. JAMA Network Open. 2020; 3(3):e201210.
[11] Dell'Osso B, Albert U, Carra G, et al. How to Improve Adherence to Antidepressant Treatments in Patients with Major Depression: A Psychoeducational Consensus Checklist. Annals of General Psychiatry. 2020; 19(61).
[12] Piette JD, Heisler M, Wagner TH. Cost-Related Medication Underuse Among Chronically Ill Adults: The Treatments People Forgo, How Often, And Who Is At Risk. American Journal of Public Health. 2004;941782-1787.
[13] Bambauer KZ, Safran DG, Ross-Degnan D, et al. Depression and Cost-Related Medication Nonadherence in Medicare Beneficiaries. JAMA Psychiatry. 2007; 64(5):602-608.
[14] Dell'Osso B, Albert U, Carra G, et al. How to Improve Adherence to Antidepressant Treatments in Patients with Major Depression: A Psychoeducational Consensus Checklist. Annals of General Psychiatry. 2020; 19(61).

disproportionately large medical expenses, one study investigated enhanced primary care depression management. This care approach involved physicians and care managers encouraging depressed patients to engage in active treatment and providing them with regularly scheduled care management during the course of a year. For patients with major depression, enhanced primary care depression management was found to have superior cost effectiveness compared to "regular treatment," demonstrating that this type of ongoing depression disease management can increase clinical improvement and be less costly over time.[15] Another source estimates that $52 billion could be saved if mental health treatments were integrated with medical treatments, in a way that supports shared responsibility among different types of providers.[16]

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 30% of the mean score (36% for TINs, 40% for TIN-NPIs). Additionally, for both TINs and TIN-NPIs, the 90th percentile score is more than twice the 10th percentile score. The distributions show meaningful variation in cost performance and suggest there's room for improvement in the costs of care.

Table 1. Distribution of the Measure Score

Metric	TIN	TIN-NPI
Count	16,208	21,802
Mean Score	$1,476	$1,429
Score Standard Deviation	$543	$539
Minimum Score	$231	$231
Maximum Score	$8,212	$8,212
Score Interquartile Range (IQR)	$532	$575
Score Percentile		
10th	$947	$897
20th	$1,090	$1,027
30th	$1,194	$1,134
40th	$1,289	$1,231
50th	$1,380	$1,333
60th	$1,481	$1,445
70th	$1,604	$1,573
80th	$1,765	$1,744
90th	$2,090	$2,038

[15] Rost K, Pyne J, Dickinson LM, LoSasso A. Cost-Effectiveness of Enhancing Primary Care Depression Management on an Ongoing Basis. Annals of Family Medicine. 2005; 3(1): 7-14.
[16] Bao Y, Casalino LP, Pincus HA. Behavioral Health and Health Care Reform Models: Patient-Centered Medical Home, Health Home, and Accountable Care Organization. Journal of Behavioral Health Services and Research. 2013; 40(1):121-132.

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 with a minimum of 20 episodes 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, Long-Term Minimum Data Set (MDS), Medicare Enrollment Database (EDB), and Common Medicare Environment (CME).

3.1.2 Specific Dataset Used for Testing

The Depression measure uses Medicare Part A and Part B, as well as Part 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 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 MDS. Specifically, the MDS is used to create the long-term care indicator variable in risk adjustment.

3.1.3 Dates of the Data Used in Testing

Depression episodes ending from January 1, 2019, through December 31, 2019.

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 characteristics of TINs and TIN-NPIs who were included in the testing of the Depression measure and attributed at least 20 episodes.

Table 2: Characteristics of Measured Entities with 20 Cases or More

Metric	TIN		TIN-NPI	
	Count	%	Count	%
Count	16,208	100%	21,802	100%
Number of Episodes Attributed	-	-	-	-
20-39 Episodes	7,714	47.6%	15,662	71.8%
40-59 Episodes	2,902	17.9%	3,809	17.5%
60-79 Episodes	1,509	9.3%	1,264	5.7%
80-99 Episodes	850	5.2%	521	2.4%
100-199 Episodes	1,719	10.6%	482	2.2%
200-299 Episodes	568	3.5%	50	0.2%
300+ Episodes	946	5.8%	14	0.06%
Census Region	-	-	-	-
Northeast	3,088	19.1%	3,863	17.7%

Metric	TIN		TIN-NPI	
	Count	%	Count	%
Midwest	3,078	19.0%	4,241	19.5%
South	7,117	43.9%	10,287	47.1%
West	2,881	17.8%	3,371	15.5%
Unknown	44	0.3%	40	0.2%

3.1.6 Patient Cohort Included in the Testing and Analysis

Table 3 shows the patient population for the Depression measure testing. It consists of Medicare beneficiaries enrolled in Medicare Parts A and B who receive care for the treatment and/or management for depression that triggers a depression episode.

Table 3: Beneficiary Demographics

Metric	Value
Count	1,769,404
Mean Age	70.4
Female %	70.6%

3.1.7 Social Risk Factors Included in Analysis

The analysis on 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.[17]

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[18]	• Variation in Medicaid eligibility across states	Yes

[17] 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

[18] Katon WJ, Lin E, Russo J, et al. Increased Medical Costs of a Population-Based Sample of Depressed Elderly Patients. JAMA Psychiatry. 2003; 60(9):897-903.

Variable	Advantages	Disadvantages	Used in Testing
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] • Only 5 categories available, which may lack granularity to fully capture disparities[20,21]	No
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., Agency for Healthcare Research and Quality (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 Depression 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 important data fields used in this measure, including diagnosis and procedure codes and other elements that are consequential to

[19] See footnote 4.
[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 & Medicaid (CMS), 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

payment. Specifically, CMS works with Zone Program Integrity Contractors, and 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 uses the Comprehensive Error Rate Testing (CERT) Program to ensure that Medicare payments are correct in accordance with coverage, coding, and billing rules. 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 a three-month claim 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 used a signal-to-noise analysis.

This approach seeks to determine the extent to which variation in the measure is due to true, underlying clinician performance, rather than random variation (i.e., statistical noise) within 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 2019, CERT estimates that proper payment, which includes payments that met Medicare coverage, coding, and billing rules, ranged from 87.3% to 96.4% of total payments each year.[23] The fiscal year 2020 Medicare fee-for-service program proper payment rate was 93.7%.[24]

[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.

Clinician-level Reliability

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	16,208	0.874	0.909	99.62%	91.59%
TIN-NPI	21,802	0.801	0.835	98.61%	79.23%

3.2.4 Interpretation

The results of the data element testing show very high reliability of the critical data elements used by the measure. The measure is highly reliable for both the TIN and TIN-NPI reporting levels, at 0.874 and 0.801 respectively. For reference, CMS generally considers 0.4 as the threshold indicating 'moderate' reliability and 0.7 as high reliability.[25] Additionally, at each testing volume threshold, the vast majority of TINs and TIN-NPI meet or exceed the moderate reliability threshold of 0.4 and most are above the high reliability threshold of 0.7.

3.3 Validity Testing

3.3.1 Level of Validity Testing

The validity of the measure was tested using face validity and empirical validity at the group/practice (TIN) and individual clinician (TIN-NPI) levels.

3.3.2 Method of Validity Testing

Face Validity

The Depression measure was developed through a structured, iterative process for gathering detailed input from recognized clinician experts on the measure. 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 good from poor performance).

In developing this measure, Acumen incorporated input from:

(i) a Depression Clinician Expert Workgroup.
(ii) a Technical Expert Panel (TEP).
(iii) the Person and Family Partners.

This process is detailed in the Episode-Based Cost Measures Development Process document posted on the MACRA Feedback Page.[26]

One of the key roles of the measure-specific Clinician Expert Workgroup is to develop service assignment rules for the cost measure. These service assignment rules are intended to ensure clinicians are evaluated on services and costs that are clinically related to the attributed

[25] 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.
[26] CMS, MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Payment-Program/Quality-Payment-Program/Give-Feedback.

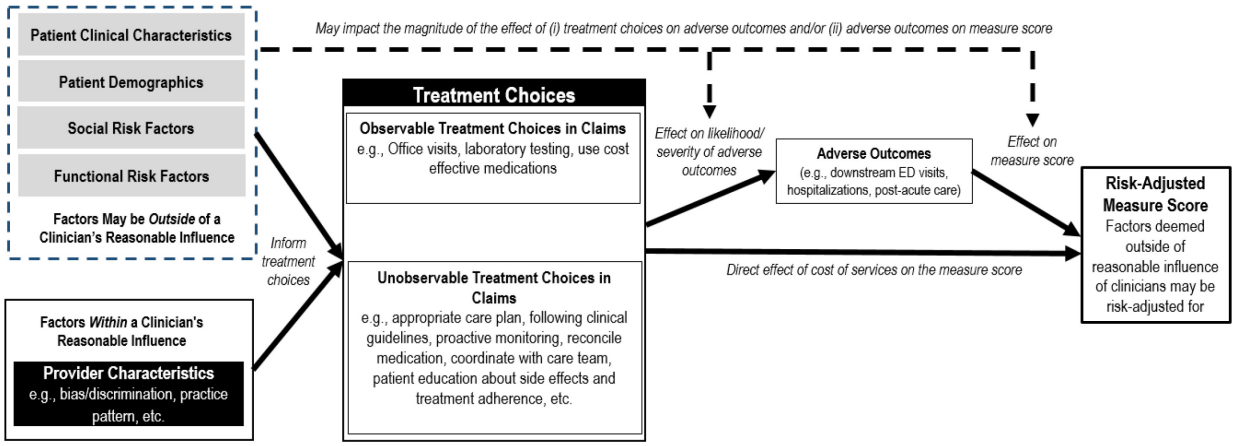
clinician's role in treating and managing the condition, thus limiting cost variation unrelated to clinician care for this measure. Therefore, assigned services are services that the Clinical Expert Workgroup believed an attributed clinician can influence via their occurrence, frequency, or intensity.

Prior to submitting the measure for the Measure Under Consideration (MUC) list, members of the Clinician Expert Workgroup were asked to consider the measure as specified and rate the degree to which the actions outlined in the logic model are within the reasonable influence of an attributed clinician, and by extension, can affect patient health outcomes and downstream costs.

Empirical Validity Testing

We evaluated the empirical validity of the Depression 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. For more information on the conceptual model, please see Section 3.5.3.

Figure 2: Conceptual Model of the Relationship between Treatment Choices and the Measure Score



The cost measure is designed to reflect the cost directly related to treatment choices, as well as the cost of adverse outcomes as a result of care. Therefore, treatment choices, either observable in claims or otherwise, by an attributed clinician can directly impact the measure score or indirectly when they're mediated through the cost of adverse outcomes. The cost of adverse outcomes, in turn, contributes to the total costs that are captured by the measure score.

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

Generally, adverse outcomes are non-trigger inpatient hospitalizations, non-trigger emergency room visits, and post-acute care. The remaining service 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

Face Validity

Figures 3 to 7 show the responses of the Clinical Expert Workgroup Members, when asked to consider the measure as specified and rate the degree to which the actions by an attributed clinician outlined in the logic model are within their reasonable influence and can affect patient health outcomes and downstream costs.

Figure 3: Reponses of Clinical Expert Workgroup Members when Asked to Rate the Degree of Influence of Attributed Clinicians over Actions Outlined in the Logic Model

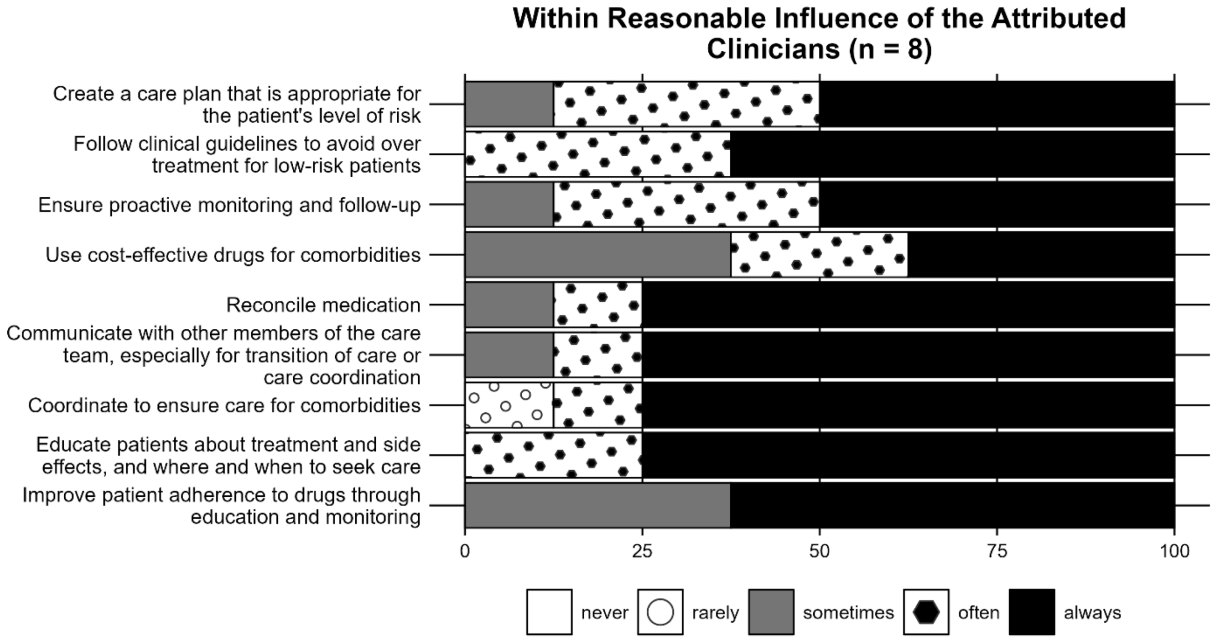


Figure 4: Reponses of Clinical Expert Workgroup Members when Asked to Rate the Likelihood of Impact on Risk of High-Cost Events for Actions Outlined in the Logic Model

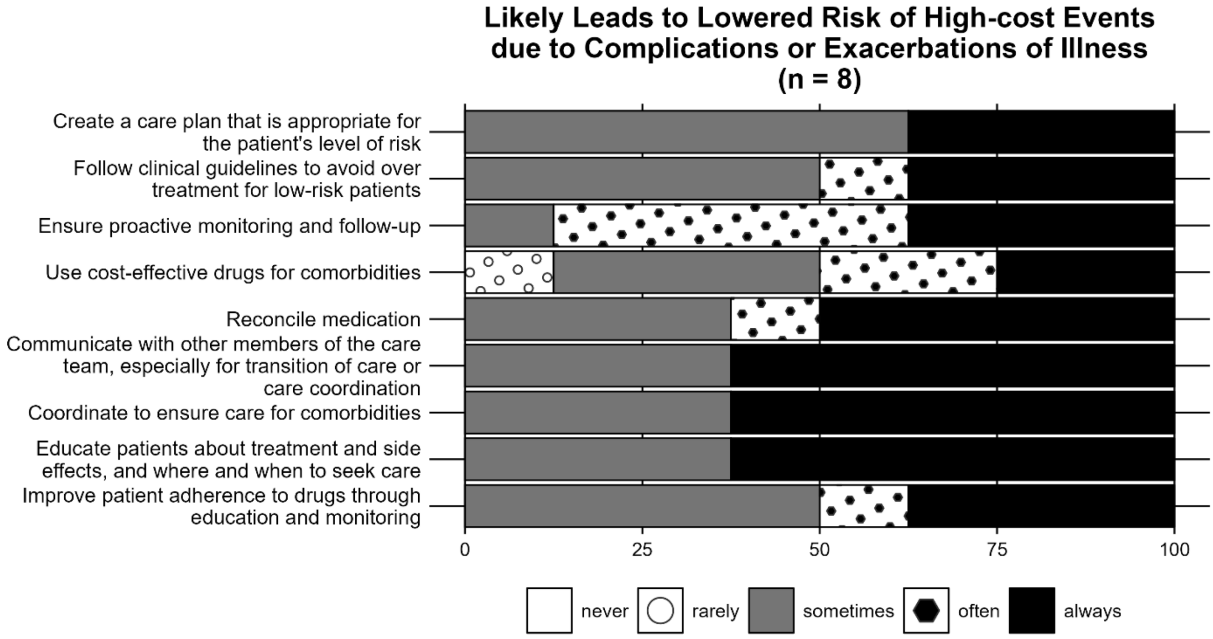


Figure 5: Reponses of Clinical Expert Workgroup Members when Asked to Rate the Likelihood of Improving Patient Treatment and Quality of Life for Actions Outlined in the Logic Model

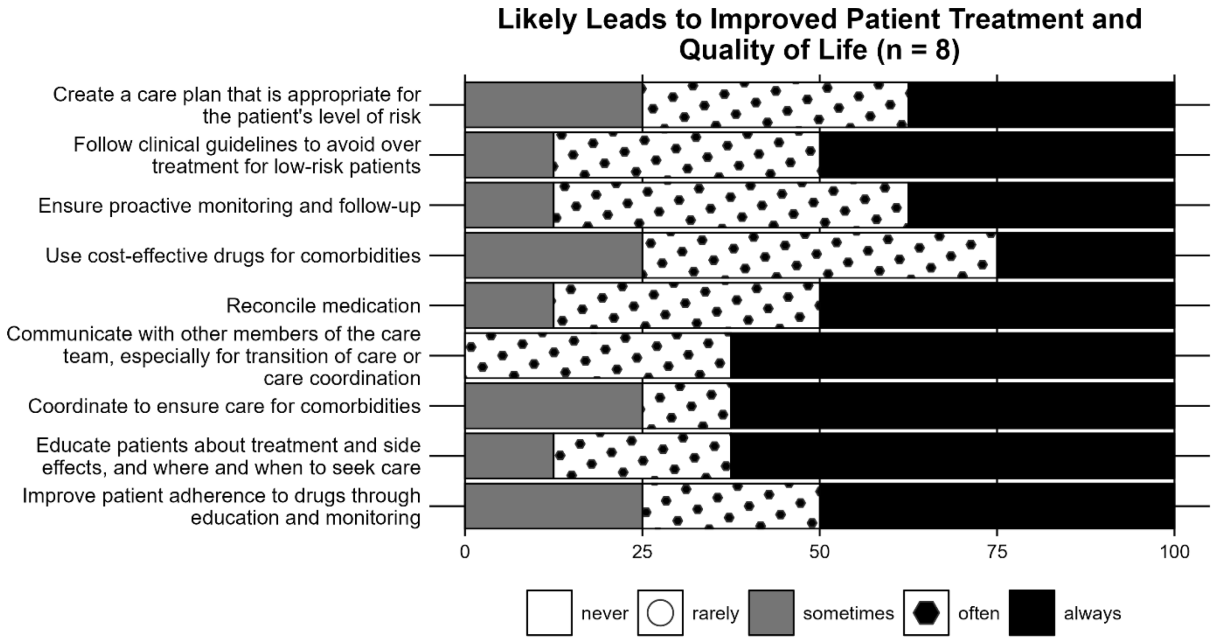


Figure 6: Reponses of Clinical Expert Workgroup Members when Asked to Rate the Likelihood of Improving Monitoring and Care Coordination for Actions Outlined in the Logic Model

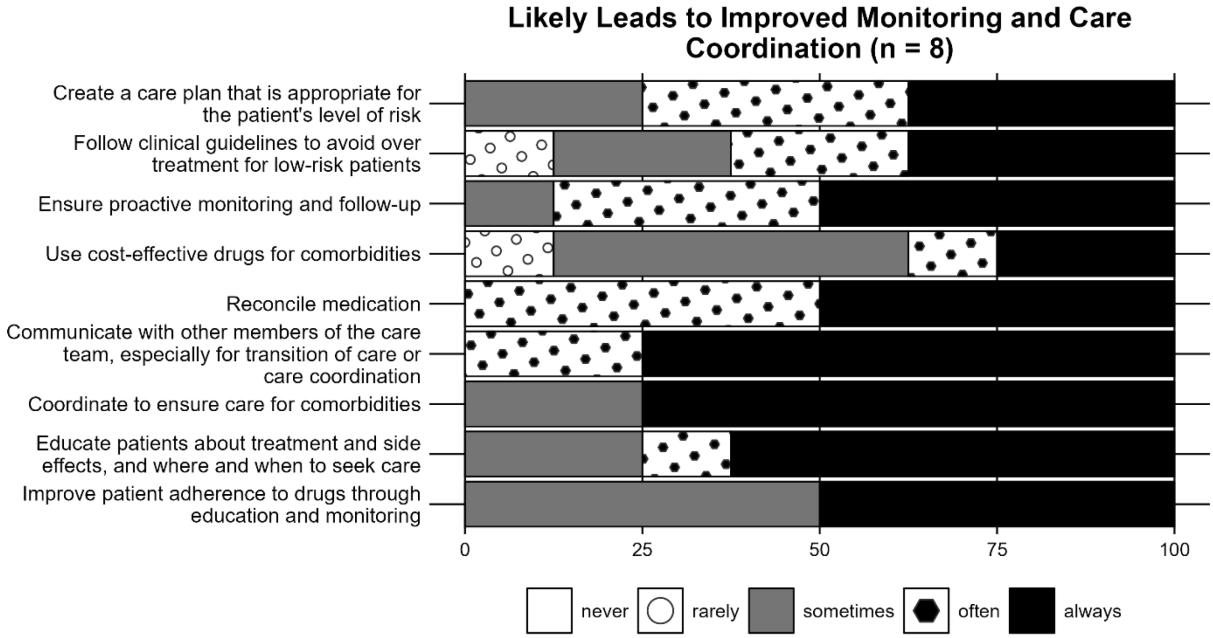
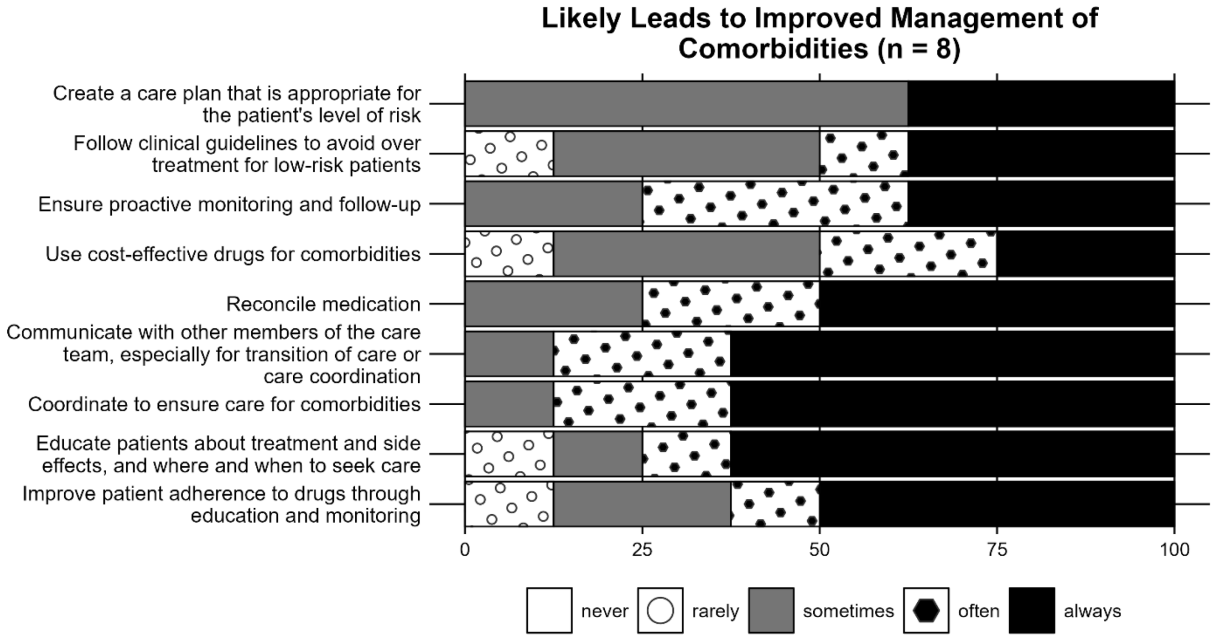


Figure 7: Reponses of Clinical Expert Workgroup Members when Asked to Rate the Likelihood of Improving Management of Comorbidities for Actions Outlined in the Logic Model



Empirical Validity Testing

Table 6 shows 2 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 $1,000 of a service category that's 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 $1,000 of a service category that's assigned to an episode, on average, while holding the remaining categories constant.

Table 6. Estimated Effect of Treatment Choices

Categories of Service	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	0.58 [0.57, 0.59] (p <0.01)	-	0.56 [0.55, 0.57] (p <0.01)	-
Outpatient Evaluation & Management Services	0.30 [0.29, 0.30] (p <0.01)	0.04 [0.03, 0.05] (p <0.01)	0.29 [0.28, 0.29] (p <0.01)	0.06 [0.05, 0.07] (p <0.01)
Laboratory, Pathology, and Other Tests	1.26 [1.17, 1.35] (p <0.01)	-0.08 [-0.22, 0.06] (p = 0.29)	1.26 [1.18, 1.34] (p <0.01)	-0.12 [-0.24, -0.00] (p = 0.04)
Chemotherapy and Other Part B-Covered Drugs	0.11 [0.09, 0.13] (p <0.01)	0.00 [-0.03, 0.04] (p = 0.83)	0.09 [0.07, 0.11] (p <0.01)	0.02 [-0.01, 0.05] (p = 0.28)
Part-D Drugs	0.07 [0.06, 0.09] (p <0.01)	-0.03 [-0.05, -0.01] (p <0.01)	0.09 [0.08, 0.10] (p <0.01)	-0.02 [-0.04, -0.00] (p = 0.02)

3.3.4 Interpretation

Face Validity

Overall, there's very strong consensus among the members that all of the actions outlined in the logic model are often or always within a reasonable influence of the attributed clinician, with every action receiving above 50% for responses that rated 'often' or 'always' (Figure 3).

When asked if the actions outlined in the logic model can influence downstream high-cost events due to complications or exacerbations of illness, with the exception of 'create a care plan that is appropriate for the patient's level of risk', all other actions received 50% or more responses that rated them as 'often' or 'always' to lead to lowered risk of downstream high-cost events if done by the attributed clinician (Figure 4). Even for 'creating a care plan that is appropriate for the patient's level of risk,' no one rated that it 'never' or 'rarely' leads to lowered risk of downstream high-cost events, while 62.5% rated it 'sometimes' and 37.5% rated it 'always'.

There's a very strong consensus among the members that all of the actions outlined in the logic model can lead to improved patient treatment and quality of life, with all actions receiving 75% or more of the responses that rated 'often' or 'always' (Figure 5).

With the exception of using cost-effective drugs for comorbidities, there's a consensus among the members that all other actions outlined in the logic model can lead to improved monitoring

and care coordination, with those actions receiving 50% or more responses that rated ‘often’ or ‘always’ (Figure 6). For using cost-effective drugs for comorbidities, 50% of the responses rated it ‘sometimes’ and 37.5% rated it ‘often’ or ‘always,’ which suggest that there’s a possibility that using cost-effective drugs for comorbidities can lead to improved monitoring and care coordination, but there may be some uncertainty for such an outcome.

With the exception of ‘creating a care plan that is appropriate for the patient’s level of risk’, all other actions outlined by the logic model were rated by 50% or more of the members to be ‘often’ or ‘always’ lead to improved management of comorbidities (Figure 7). For creating a care plan that is appropriate for the patient’s level of risk, no one rated rarely or never, all members rated sometime or always, which suggests that while such outcome is uncertain, there is a possibility that the action may be effective.

Empirical Validity Testing

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.

The results are also consistent with performance gaps identified from the literature review in Section 2.2.1, such as potentially avoidable hospitalization and emergency department visits. Model 1 shows that the cost of adverse events is associated with a worse measure score, which includes hospitalizations or emergency department visits that are clinically related to depression. The measure score is shown to be increasing with outpatient evaluation and management services, laboratory services, and Part B and Part D drugs. However, laboratory services and Part D drugs appear to also influence the measure score by decreasing the cost of adverse events as shown in model 2. This pattern suggests that, while these treatment choices are able to reduce the risk of adverse events and help improve the measure score, they may also be prone to overuse.

Outpatient evaluation and management services are associated with a worse score and higher cost of adverse outcomes, which may reflect higher service intensity that are linked to adverse outcomes and overall higher usage among depression patients, as suggested by the literature. On the other hand, Part B drugs are shown to be associated with a worse score, but not with adverse events, which may indicate that their cost is mostly captured directly by the measure score.

3.4 Exclusions Analysis

3.4.1 Method of Testing Exclusions

Exclusions are used in the Depression measure to ensure:

- a comparable patient population within the scope of the measure's focus on a clinician's or clinician group's risk-adjusted cost to Medicare for patients receiving medical care to manage and treat depression.
- that episodes provide meaningful information to attributed clinicians.
- that sufficient data (as part of data processing) 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 that are shorter than one year
- Episodes where patient death date occurred before the episode end date
- Outlier episodes that can't be reliably predicted by the risk adjustment model
- Episodes where there isn't an attributed clinician
- Presence of Bipolar Disorder Pre- and Post-Trigger
- Presence of Schizophrenia Pre- and Post-Trigger
- Presence of Drug/Alcohol Psychosis Pre- and Post-Trigger
- Episodes where the attributed clinicians haven't reached a minimum of 20 episodes

Given the rationales for these exclusions, we would 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 measure calculation to assess the distinctness between the 2 patient cohorts. A full list of the exclusions used for the Depression measure is provided in the Measure Codes List available on the MACRA Feedback Page.[27]

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 the TIN and TIN-NPI levels. These exclusion criteria ensure that the reportable episode populations are more homogenous and comparable than all episodes meeting the triggering logic. It's worth noting that only the observed cost is shown, which hasn't been risk-adjusted using our risk adjustment model. Therefore, the differences in cost may appear much smaller after risk adjustment than as-is.

[27]CMS, MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Payment-Program/Quality-Payment-Program/Give-Feedback.

Table 7: Cost Statistics for Measure Exclusions

Exclusion	Episodes		Observed Cost					
			Mean	Percentile				
	#	%		10th	25th	50th	75th	90th
All Episodes Meeting Triggering Logic	2,452,137	100.00%	$1,920	$215	$375	$803	$2,094	$4,693
Episode Length Less Than One Year	92,202	3.76%	$4,060	$523	$941	$2,007	$4,274	$8,471
Beneficiary Death in Episode	174,234	7.11%	$3,517	$416	$783	$1,760	$3,802	$7,505
Outlier	40,326	1.64%	$5,102	$190	$426	$6,213	$9,882	$9,882
No Attributed NPI (TIN-NPI Reporting Only)	116,739	4.76%	$2,562	$337	$594	$1,305	$2,971	$6,099
Presence of Bipolar Disorder Pre-Trigger	115,327	4.70%	$3,976	$365	$754	$1,732	$4,386	$10,273
Presence of Bipolar Disorder Post-Trigger	190,438	7.77%	$3,971	$371	$762	$1,753	$4,413	$10,083
Presence of Drug/Alcohol Psychosis Pre-Trigger	6,654	0.27%	$3,828	$315	$702	$1,706	$4,055	$8,866
Presence of Drug/Alcohol Psychosis Post-Trigger	13,128	0.54%	$4,255	$377	$823	$1,994	$4,745	$10,215
Presence of Schizophrenia Pre-Trigger	61,130	2.49%	$5,543	$464	$979	$2,299	$6,352	$15,026
Presence of Schizophrenia Post-Trigger	92,152	3.76%	$5,371	$462	$972	$2,289	$6,103	$14,108
TIN doesn't Meet Case Minimum	436,798	17.81%	$2,219	$205	$393	$988	$2,632	$5,491
TIN-NPI doesn't Meet Case Minimum	1,336,497	54.50%	$1,934	$201	$350	$776	$2,119	$4,756
Reportable Episodes (if all clinicians reported as TIN at the testing case minimum)	1,625,048	66.27%	$1,381	$202	$339	$655	$1,581	$3,468
Reportable Episodes (if all clinicians reported as TIN-NPI at the testing case minimum)	808,630	32.98%	$1,400	$215	$360	$683	$1,586	$3,490

3.4.3 Interpretation

Overall, exclusion criteria decrease the distribution of observed cost of all episodes meeting trigger logic, from the mean of $1,920 to $1,381 at the TIN reporting level and $1,400 at the TIN-NPI reporting level (Table 7). All of the exclusion criteria have higher mean observed cost than all episodes meeting triggering logic.

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're likely to be included in the following performance period once the episode length is longer than one year.

Episodes where a patient died before the episode end date are excluded because they don't provide sufficient data in the episode window period. These episodes also have a higher mean observed cost than all episodes meeting the triggering logic, at $3,517, 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 $5,102 compared to $1,920 for all episodes meeting the triggering logic. The wide variability of observed episode costs for outlier cases also supports their exclusion.

Episodes where there isn't an attributed clinician are excluded because these episodes don't have any TIN-NPIs that billed at least 30% of the clinically-related claims with a relevant diagnosis**.** Failing to meet the attribution rules indicates that a provider hasn't assumed a significant role in the care of the patient or the patient-clinician relationship. Their mean observed cost, at $2,562, is higher than all episodes meeting the triggering logic, at $1,920.

Based on the input from the clinical expert workgroup, several comorbidities are excluded because these episodes can be clinically distinct from the overall major depressive disorder population. Specifically, the presence of bipolar disorder, drug/alcohol psychosis, and schizophrenia both before the episode's trigger and during the episode are exclusion criteria recommended by the workgroup. These episodes have mean observed costs that are at least 2 times higher than all episodes meeting the trigger logic, which suggests that they may have distinct resource use patterns from a typical episode.

The largest exclusions come from applying the case minimum requirements, to ensure that low-volume providers aren't disadvantaged. This is because their scores are prone to disproportional swings due to outlying events or random noise. The mean observed cost of these episodes is higher than all episodes meeting the triggering logic, which may suggest that economy of scale can play a role in controlling costs.

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 162 risk factors and stratification by 2 risk categories.

The risk adjustment model for the Depression 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
- dual eligibility status

The model also includes measure-specific factors:

- Chronic Pain
- Eating Disorder

- Memory Loss
- Suicide Attempt
- Suicide Ideation
- 2 or more prior hospitalizations in one year
- Any prior observational care in the lookback window
- Prior ECT within one year before episode of care
- Prior Esketamine within one year before episode of care
- Prior TMS within one year before episode of care

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

- Depression with Psychotic Features
- Depression without Psychotic Features
- Depression with Psychotic Features & Medicare Part D enrollment
- Depression without Psychotic Features & Medicare Part D enrollment

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 MACRA Feedback page.[28]

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 Depression 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 in Section 3.3.2 shows the conceptual model that outlines how SRFs can influence the measure score, which is informed by both published external research and our own data

[28]CMS, MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Payment-Program/Quality-Payment-Program/Give-Feedback.

analysis.[29,30,31,32,33] 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 of the influence of the attributed clinician. Risk factors, including SRFs, can both influence the treatment choices and impact the size of the effect of treatment choices on mitigating the risk of adverse outcomes and the cost of adverse outcomes.

A systematic approach then guides the decision of which factors to include in the risk adjustment model. First, we reviewed the literature to gather known risk factors and drivers of resource use. These factors are usually diagnoses; therefore, the first set of risk adjustors are commonly the HCCs. Then, we consulted our clinical expert panels on additional factors that are known to be associated with resource use. Together with our clinical expert panel, we reviewed the stratified results on episode cost across many different patient characteristics. We arrived at the final list of risk adjustors based on those discussions and consensus among the clinical experts. Additionally, during our testing phases, we also follow a structured and systematic approach to decide whether SRFs should be adjusted for, which is further described in Section 3.5.5.

3.5.4 Statistical Results

The literature has extensively tested the use of the HCC model as applied to Medicare claims data. Although the variables in the HCC model were chosen to predict annual cost, CMS has also used this risk adjustment model in a number of 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 the 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 V22 2016 model can be found in the Evaluation of the CMS-HCC Risk-Adjustment Model report[34] and the Report to Congress: Risk Adjustment in Medicare Advantage.[35] 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 measure sub-groups.

3.5.5 Analyses and Interpretation in Selection of Social Risk Factors

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

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

---

[29] See Footnote 15.
[30]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.
[31]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
[32]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/.
[33] 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-based-purchasing-programs
[34]Pope, Gregory C., John Kautter, et al., "Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report." RTI International: March 2011.
[35]CMS, "Report to Congress: Risk Adjustment in Medicare Advantage," https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/RTC-Dec2018.pdf.

(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 the clinician's fixed effects,

(iii) whether the 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 doesn't risk adjust for SRFs.

Overall, the results suggest that it's appropriate to risk adjust for social risk factors in this measure. There's a statistically significant association between the patient's dual status and episode cost, as observed on the largest subgroups (Table 8). This association is relatively stable in the largest subgroups and becoming statistically significant after adding variables to account for provider-level factors, which suggests that the patient-level factors are more influential than provider-level factors. This is also supported by the evidence that the performance degradation is observed mainly on dual episodes (Table 9). While many providers are able to perform equally well on their dual episodes and non-dual episodes, there are many more providers who are performing significantly worse on their dual episodes than their non-dual episodes, which suggests that providers aren't able to fully mitigate the effect of SRFs (Table 10). Lastly, risk adjusting for dual status appears to substantially change the performance ranking for many providers (Table 11).

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

Level	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	Depression with Psychotic Feature with Part D Coverage	2.54%	$68.98 (p = 0.17)	$347.67 (p < 0.01)	$272.92 (p < 0.01)
TIN	Depression with Psychotic Feature without Part D Coverage	0.54%	$430.57 (p = 0.09)	$ 500.09 (p = 0.05)	$565.58 (p = 0.12)
TIN	Depression without Psychotic Features with Part D Coverage	74.98%	$254.10 (p < 0.01)	$231.14 (p < 0.01)	$192.48 (p < 0.01)
TIN	Depression without Psychotic Features without Part D Coverage	21.94%	$120.22 (p < 0.01)	$123.93 (p < 0.01)	$131.94 (p < 0.01)
TIN-NPI	Depression with Psychotic Feature with Part D Coverage	2.50%	$97.8 (p = 0.06)	$ 406.72 (p < 0.01)	$376.11 (p < 0.01)
TIN-NPI	Depression with Psychotic Feature without Part D Coverage	0.53%	$513.88 (p = 0.05)	$637.66 (p = 0.02)	$864.75 (p = 0.24)
TIN-NPI	Depression without Psychotic Features with Part D Coverage	75.0%	$252.00 (p < 0.01)	$238.04 (p < 0.01)	$202.43 (p < 0.01)
TIN-NPI	Depression without Psychotic Features without Part D Coverage	21.97%	$110.07 (p < 0.01)	$114.25 (p < 0.01)	$173.12 (p < 0.01)

Table 9: Mean Ratio of 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	1.03	1.07	1.01	1.01	1.05	0.99
0%	1.01	-	1.01	0.99	-	0.99
1-20%	1.00	1.05	1.00	0.98	1.02	0.97
21-40%	1.04	1.08	1.02	1.01	1.06	0.99
41-60%	1.07	1.11	1.03	1.06	1.10	1.01
61-80%	1.06	1.07	1.04	1.06	1.08	1.01
81-99%	1.15	1.16	1.01	1.15	1.16	1.02
100%	1.15	1.15	-	1.13	1.13	-

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

Reporting Level	Significantly Worse	Equally Well	Significantly Better
TIN	7.96%	89.98%	2.06%
TIN-NPI	7.01%	91.49%	1.5%

Table 11. Clinicians' Performance Shift Measured by the Change in the Average Ratio of Observed-to-Expected Cost

Reporting Level	Proportion of Clinicians Affected at Various Levels of Performance Shift	
	Ranking Shift by 1% or more	Ranking Shift by 5% or more
TIN	76.2%	8.7%
TIN-NPI	74.7%	7.5%

3.5.6 Method for Statistical Model or Stratification Development

To analyze the validity of current risk adjustment model, we examined 2 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 the 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 Depression cost measure, calculated by dividing the explained sum of squares by the total sum of squares, is 0.18. The adjusted R-squared is 0.18. More information on discrimination testing for the CMS-HCC model is available at Pope et al. 2011.[36]

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 0.95 to 1.03 across all risk deciles (with an overall average of 1.0).

Table 12: Predictive Ratio by Decile of Predicted Episode Cost

Decile	Average Predictive Ratio
Decile 1	0.95
Decile 2	0.99
Decile 3	0.99
Decile 4	0.99
Decile 5	1.01
Decile 6	1.00
Decile 7	1.03
Decile 8	1.02
Decile 9	1.00
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 similar to or higher than the values presented in similar analyses of risk adjustment models.[37] 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 aren't 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 isn't essential because not all of the variation in the cost of care should be adjusted. In collaboration with the experts from our clinical workgroup, this measure only adjusts for factors that are deemed to be outside of the influence of clinicians.

Table 12 shows 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 0.95 and 1.03. Overall, the risk adjustment model doesn't over- or under-predict cost across the full range of resource use patterns in the population.

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

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 high-cost 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. In addition, the testing results found that 8.5% of episodes had at least one clinically related emergency department visit with a mean risk-adjusted episode cost of $2,971 and 0.3% of episodes had at least one clinically related acute inpatient stay with a mean episode cost of $5,588.

3.6.3 Interpretation

The results suggest that there's opportunity for improvement in performance across providers. There's 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 (Table 1). The measure score at the 90th percentile is over 2 times greater than the measure score at the 10th percentile at both the TIN and TIN-NPI levels. There are also opportunities to reduce costs associated with high-cost events, such as clinically related emergency department visits and acute inpatient stays. Episodes with a clinically related emergency department visit cost Medicare approximately $264 million more than an average Depression episode, and $28 million for episodes with a clinically related acute inpatient stay.

3.7 Missing Data Analysis and Minimizing Bias

3.7.1 Method

Since CMS uses Medicare claims data to calculate the Depression measure, Acumen expects a high degree of data completeness. To further ensure that we have complete and accurate data for each patient, Acumen typically excludes episodes where the patient's date of birth information (an input to the risk adjustment model) can't be found in the enrollment database, the patient's information doesn't appear in the enrollment database, the patient resides outside of the U.S., death occurred before the episode, or the primary payer isn't Medicare.

The Depression 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 patient's complete clinical profile 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

Table 13 presents the frequency and observed episode cost for categories of missing data, which caused episodes to be excluded from the Depression measure. Frequency is presented in terms of the number of episodes excluded due to missing data, as well as the cost profile. It's worth noting that only the observed cost is shown, which hasn't been risk-adjusted for using our risk adjustment model. Therefore, the differences in cost may appear much smaller after risk adjustment than as-is.

As a note, the episode counts below reflect exclusions from the initial population of triggered episodes. After the missing data exclusions are applied, we then 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	Observed Cost						
	Episode Count	Percentile					
		Mean	10th	25th	50th	75th	90th
All Episodes	3,184,241	$1,843	$187	$343	$746	$1,972	$4,480
Beneficiary Not Found in Enrollment Database	*	*	*	*	*	*	*
Beneficiary Resides Outside of U.S. or Territories	7,302	$1,533	$205	$327	$684	$1,630	$3,547
Primary Payer Other than Medicare	395,555	$1,803	$166	$314	$693	$1,840	$4,200
No Continuous Enrollment in Medicare Parts A and B, and Any Enrollment in Part C	410,157	$1,302	$104	$194	$437	$1,216	$3,085

* Cells suppressed due to having fewer than 10 observations

3.7.3 Interpretation

The results show that the missing data episodes don't appear to be substantially different than all episodes in the initial population in terms of cost (Table 13). Given their limited frequencies and minimal difference in cost profile, the impact of removing these episodes on the overall measure should be minimal while ensuring that clinicians are fairly evaluated on episodes with complete data.

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 3 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 Common Working File (CWF) maintained at the CMS Baltimore Data Center. Medicare claims are submitted by healthcare providers to a Medicare Administrative Contractor (MAC), and 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 a result, it isn't practical to wait until all claims for a given month are finalized before calculating this measure. As such, there's a trade-off between efficiency (accessing the data in a timely manner) and accuracy (waiting until most claims are finalized) when determining the length of the time (i.e., the "claims run-out" period) after which to pull claims data. To determine the appropriate claims run-out period, Acumen has performed testing on the delay between claim service dates and claims data finalization. Based on this analysis, Acumen uses a run-out period of 3 months after the end of the calendar year to collect data for development and testing purposes. If this measure is used in a CMS program, calculation and reporting would be done in line with that program's reporting practices.

4.3.1.2 Missing Data

This measure requires complete beneficiary information, and a small number of episodes with missing data are excluded to ensure completeness of data and accurate comparability across episodes. For example, episodes where the beneficiary wasn't enrolled in Medicare Parts A and B for the 120 days prior to the episode start date aren't included in this measure. This 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 can't be located aren't included in this measure.

4.3.1.3 Sampling

During measure testing, Acumen noted that episodes in which the beneficiary died prior to the episode end date exhibited different cost distributions compared to other episodes. To avoid this effect's potential impact on clinician scores, this measure doesn't include episodes for which the beneficiary's date of death occurs prior to the end of the episode window.

5.0 Usability and Use

5.1 Use

5.1.1 Current and Planned Use

The Depression measure isn't currently in use, but it's intended for use in a payment program and could eventually be publicly reported. The measure 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 the measure to be used in MIPS, it must be reviewed by the Measure Application Partnership (MAP) and then undergo the notice-and-rulemaking process. Given these next steps, the earliest the measure could be in use in MIPS is calendar year 2024. 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 Depression measure development, we used an iterative and extensive process to gather feedback on the measure and its results to ensure that the measure can be used appropriately in the MIPS program by clinicians and clinician groups who practice in this clinical area. This process also aims to make sure that the measure performance results can be understood by the population that's being measured, to help support decision making. A couple of the main ways that we gathered feedback was through i) reoccurring Clinician Expert Workgroup meetings, where members discussed the clinical perspective, the patient perspective, and empirical data, in order to recommend measure specifications, and ii) the national field testing of the measure.

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). 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 was appropriately assessing costs for the attributed clinicians.

Field Testing

Additionally, Acumen and CMS nationally field tested the draft Depression measure, along with 4 other episode-based cost measures, for a 10-week comment period (January 10 to March 25, 2022). We provided a Field Test Report with performance data to all clinician groups and clinicians who were attributed 20 or more episodes.[38] 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 clinicians and other interested members of the public as possible. A total of 17,237 TIN reports and 23,927 TIN-NPI 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.

[38]The field test reports are 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 in advance of or during each of the following Clinician Expert Workgroup Meetings:

- Workgroup meeting
- Service Assignment and Refinement Meeting
- Post-Field Test Refinement Meeting.

During the meetings, Acumen guided 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 was appropriately assessing 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, feedback on the appropriateness of the measures and the usability of the data was gathered from clinician and clinician groups who received a report as well as from the general public. Comments from field testing were summarized in a public report, which was also shared with the Clinician Expert Workgroup to consider in 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.

5.1.2.2.1 Data Provided During Field Testing

Each Field Test Report contained:

- Detailed performance results for the attributed measure, including the 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)[39]
    - 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, and the patient relationship code (PRC) on the trigger/reaffirming claim line.

All interested members of the public, including those who didn't 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

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

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, a Measure Testing Form (including reliability and validity data), and a National Summary Data Report (including national level summary statistics on the measure).[40] During field testing, Acumen conducted education and outreach activities including multiple office hours sessions with specialty societies, a publicly posted field testing webinar recording, and the Quality Payment Program Help Desk support.

5.1.2.2.2 Education and Outreach

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

Acumen and CMS hosted 2 office hours sessions in January 2022 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 over 35 attendees from targeted specialty societies who are likely to have members who could be attributed the measure.

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

Acumen and CMS posted the MACRA Wave 4 Cost Measures Field Testing Webinar to the Quality Payment Program Webinar Library at the start of the field testing period.[41] The webinar recording, slides, and transcript were publicly available for review throughout field testing. The webinar presentation outlined: (i) the cost measure field testing project (ii) the measure development and re-evaluation processes, and (iii) field testing activities.

5.1.2.3 Feedback on Measure Performance and Implementation

Clinician Expert Workgroup Meetings

Feedback from the Workgroup members was recorded throughout the meeting. More formal feedback was gathered using polls, typically requesting 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.

[40]The measure specifications, mock reports, Measure Development Process document, Frequently Asked Questions document, and testing documents are posted on the MACRA Feedback Page: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/MACRA-Feedback.html.

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

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.[42] The 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 Depression measure.

5.1.2.5 Feedback from Other Users

Person and Family Engagement

Acumen incorporated input from patients and caregivers throughout the Depression measure development process. Before each Clinical Expert Workgroup meeting, Person and Family Partners (PFPs) provided input through focus groups and interviews to help inform the Workgroup's discussion. Attending PFPs then presented the findings for the Workgroup members, which helped shape the recommendations they made for the measure specifications. Some examples of feedback the PFP gave include improving the care coordination between primary care providers and specialists (e.g. psychologists and psychiatrists) and eliminating barriers to access to consistent care, such as lack of insurance or available professionals. With consideration of the PFP findings, the Depression measure includes telehealth codes as condition-related Current Procedural Terminology / Healthcare Common Procedure Coding System (CPT/HCPCS) codes.

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, which was comprised of subject matter and measure development experts. Acumen conducted analyses into potential adjustments that could be made to the measures to improve the measures' ability to assess the intended clinician population.

After completing field testing, Acumen compiled the feedback provided through the survey 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 Depression measure made after consideration of field testing analyses and feedback are:

- The name of the measure was changed from "Major Depressive Disorder" to "Depression" to better reflect the scope of the measure.
- Acumen removed nursing facility E&M codes 99304–99310, 99315–99316, and 99318 from the measure's trigger logic.
- Acumen added additional codes (CPT/HCPCS 96101, 96130, 96132, 96136) related to psychological and neurological testing to aid in triggering and confirming episodes.
- Acumen added a measure-specific risk adjustor variable for observation stays as an indicator of TRD.
- The following classes of drugs were removed from the measure:

[42]CMS, "2020 Field Testing Feedback Summary Report," MACRA Feedback Page, https://www.cms.gov/files/document/macra-2020-ft-feedback-summary-report.pdf.

    - Antihistamines – Piperidines Movement Disorder Drug Therapy
    - N-Methyl-D-aspartic acid (NMDA)
    - Vasomotor Symptom Agents
- Acumen removed ICD-10 codes G21, G24, G26, and F53 from the measure's service assignment rules.

5.2 Usability

5.2.1 Improvement

The measure hasn't yet been implemented, and as such hasn't had influence over performance. Our testing suggests that there's 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 assesses 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 hasn't been implemented at this time, so we don't 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 more treatment services may be associated with a worse score, it's often mediated by the cost of adverse events. In other words, attempting to stint on care will lead to an increased risk of downstream adverse events that will in turn be detrimental to the cost measure score. Therefore, it isn't in a clinician's best interest to do so to optimize their score.

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 3 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 hasn't been implemented at this time, there are no testing results that identify unexpected benefits. However, currently, many clinicians can only be assessed by the MSPB-Clinician and TPCC measures in the cost performance category. 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 Depression Episode Group

Measure Title	Measure ID	Measure Description	Measure Type
Anti-Depressant Medication Management	Q009	Percentage of patients 18 years of age and older who were treated with antidepressant medication, had a diagnosis of major depression, and who remained on an antidepressant medication treatment. Two rates are reported. a. Percentage of patients who remained on an antidepressant medication for at least 84 days (12 weeks). b. Percentage of patients who remained on an antidepressant medication for at least 180 days (6 months).	Process
Adult Major Depressive Disorder (MDD): Suicide Risk Assessment	Q 107	Percentage of patients aged 18 years and older with a diagnosis of major depressive disorder (MDD) with a suicide risk assessment completed during the visit in which a new diagnosis or recurrent episode was identified.	Process
Preventive Care and Screening: Screening for Depression and Follow-Up Plan	Q134	Percentage of patients aged 12 years and older screened for depression on the date of the encounter or 14 days prior to the date of the encounter using an age appropriate standardized depression screening tool AND if positive, a follow-up plan is documented on the date of the eligible encounter.	Process
Depression Remission at Twelve Months	Q 370	The percentage of adolescent patients 12 to 17 years of age and adult patients 18 years of age or older with major depression or dysthymia who reached remission 12 months (+/- 60 days) after an index event date.	Outcome

The MIPS quality measures listed above are related to the depression measure as they directly manage, treat, and monitor the aforementioned condition. These quality measures include 1 outcome measure on depression remission (Q370) and 3 process measures on medication management (Q009), suicide risk assessment (Q107), and screening (Q134). Most of these quality measures are specific to depression, with one that applies to a broader cohort of patients with depression in general. All of these quality measures are valuable in ensuring that clinicians are measured on both cost and quality for this episode group.

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). No consensus was reached.

6.3 Competing Measures

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

Additional Information

Depression 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.

Barbara Spivak, MD, Mount Auburn Cambridge Independent Practice Association (MACIPA)
Becky Fenton, PsyD, NYC Department of Homeless Services, New York City
Carolyn Dueñas, MBA, RN, NGALE
David Kroll, MD, Brigham and Women's Hospital/Harvard University
Gerard Hogan, DNSc, CRNA, ARNP-BC
James Gajewski, MD, Veterans Administration
Jamieson Wilcox, OTD, OTR/L, University of Southern California, Los Angeles
Kate Lichtenberg, DO, MPH, FAAFP, FACPM, Anthem Blue Cross Blue Shield
Luisa Collins, MSN, FNP-C, APRN, ABAAHP, CPHIMS, Redefined Medicine
Megan Adamson, MD, MHS-CL, FAAFP, Clinica Family Health
Naakesh Dewan, MD, Florida Blue
Robert Roca, MD, Johns Hopkins University School of Medicine
Terry Lee Mills, MD, MMM, CPE, FAAFP, CommunityCare
Vaile Wright, PhD, American Psychological Association

Measure Developer Updates and Ongoing Maintenance
The measure isn't currently in use, but the earliest possible release of the measure in MIPS would be calendar year 2025. If the measure becomes finalized for use in MIPS, it would undergo annual maintenance and a comprehensive re-evaluation every 3 years. This measure has been submitted to the 2022 MUC List and may be reviewed by the MAP in the winter 2022. There are no further updates or reviews for this measure scheduled at this time.