Quality Payment
PROGRAM

Major Depressive Disorder

Measure Testing Form

January 2022

Contents

1.0 Introduction

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

1.1 Project Title and Overview

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

1.2 Measure Name

Major Depressive Disorder

1.3 Type of Measure

Cost/Resource Use

1.4 Data

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

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

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

2.0 Preliminary Testing Results

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

2.1 Measure Coverage

Table 1 shows the number of patients and the amount of Medicare Parts A and B cost covered by this measure. This measure has the potential to have high impact as major depressive disorder is a common condition for the Medicare population, as shown in the results below.

Table 1. Beneficiary and Cost Coverage

Coverage Metrics	Coverage at Volume Thresholds	
	1 Episodes	20 Episodes
Number of Patients	1,771,426	1,508,510
Percent of Parts A & B Costs Covered by the Measure – TIN	0.5%	0.4%
Percent of Parts A & B Costs Covered by the Measure – TIN-NPI	0.5%	0.2%

2.2 Frequently Attributed Specialties

Table 2 shows the top 10 attributed specialties for this measure, using a 20-episode testing volume threshold. The most frequently attributed specialties reflect the intent of the measure to capture costs of the management of major depressive disorder, including both primary care clinicians, specialists, and clinicians focusing on mental and behavioral health. These clinicians are also consistent with input provided by stakeholders, including patient and family partners (PFPs), during the measure development process. PFPs identified psychiatrists, primary care clinicians, counselors, licensed clinical social workers, and advanced practice registered nurses amongst others as being part of their care team.

Table 2 also provides metrics to show the breakdown of the share of episodes covered by each specialty. Internal medicine and family practice together make up around half of all clinicians who meet the testing volume threshold (26.4% and 22.8%, respectively). Psychiatrists are the third most frequently attributed specialty, comprising 19.2% of all attributed clinicians.

Finally, Table 2 shows the percentage of each specialty that is covered by this measure. 13.1% of all psychiatrists who billed at least one Part B physician/supplier claim in 2019 have at least 20 Major Depressive Disorder episodes. This metric should be interpreted with caution as specialties vary in size; for instance, the share of internal medicine clinicians covered by the measure appears low (4.5%) but it is a very large specialty.

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

Rank	Specialty	Number of TIN-NPIs Attributed	Percent of Specialty Among All Attributed TIN-NPIs	Percent of All Episodes	Percent of Specialty (TIN-NPI) Attributed to Measure
1	Internal Medicine	6,322	26.4%	25.1%	4.5%
2	Family Practice	5,460	22.8%	19.7%	4.8%
3	Psychiatry	4,588	19.2%	23.9%	13.1%
4	Nurse Practitioner	2,865	12.0%	12.1%	1.5%
5	Clinical Psychologist	1,804	7.5%	7.7%	5.6%
6	Licensed Clinical Social Worker	865	3.6%	2.7%	1.7%
7	Physician Assistant	442	1.8%	1.8%	0.4%
8	Neurology	384	1.6%	1.6%	1.9%
9	Geriatric Medicine	313	1.3%	1.1%	11.8%
10	General Practice	236	1.0%	1.0%	2.9%

2.3 Reliability

Reliability evaluates a measure's ability to consistently differentiate the performance of one clinician from another. The signal-to-noise ratio is used to estimate reliability, which indicates how much of the variation in the measure score is explained by differences among clinicians performance (i.e., signal) instead of differences within each clinician's performance (i.e., noise). Specifically, noise is the variation from one episode to another during the performance period for a particular clinician.

Table 3 shows reliability metrics at various testing volume thresholds. While higher thresholds yield higher reliability results, it is at the cost of further reducing the number of clinicians and clinician groups eligible for the measure, which would reduce the potential impact of the measure. For the purposes of field testing, we used a 20-episode volume threshold (bolded in the table below); for simplicity, we use this threshold across all measures. If the measure is implemented in the Merit-based Incentive Payment System (MIPS) in the future, CMS will establish a case minimum through notice-and-comment rulemaking.

Table 3. Sample Size, Mean Reliability, and Proportion of Clinicians above Moderate Reliability at Various Testing Volume Thresholds

Testing Volume Threshold	TIN			TIN-NPI		
	Number of TINs	Mean Reliability	Percent Above 0.4	Number TIN-NPIs	Mean Reliability	Percent Above 0.4
10	29,684	0.87	99.2%	60,127	0.80	97.3%
20	**17,237**	**0.92**	**100.0%**	**23,927**	**0.87**	**99.9%**
30	12,074	0.95	100.0%	12,091	0.90	100.0%

At the testing volume of 20 episodes, the mean reliability for the Major Depressive Disorder measure is high, specifically 0.92 at the TIN level and 0.87 at the TIN-NPI level (Table 3). CMS generally considers 0.4 as the threshold indicating 'moderate' reliability and 0.7 indicating 'high' reliability, which is supported by previous work into reliability and the threshold was finalized in

the 2022 Physician Fee Schedule final rule.[2,3] Almost all TINs and TIN-NPIs meet or exceed the moderate reliability threshold of 0.4 at the 20-episode testing volume threshold, 99.9% at both reporting levels.

2.4 Validity

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

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

Table 4 shows that downstream high-cost events, such as emergency department visits or acute inpatient stay, are correlated with the cost measure. Specifically, providers with higher frequency of high-cost events, indicated by the quartiles of the high cost event frequency, have higher mean scores at both the TIN and TIN-NPI reporting levels (Table 4). Providers with the highest frequency of high-cost events, at Q3 or Q4, also have mean scores that are higher than the overall mean score for all providers that indicate lower-than-average performance.

Although episodes with high-cost events are more costly than ones without, the results show that the measure score is not impacted until a provider has substantially more high-cost events than their peers. Specifically, providers with the lowest frequency of high-cost events, either 0%, Q1, or Q2 have lower mean score than the overall mean score for all providers (Table 4).

However, there are some exceptions, specifically the Q2 mean score for acute inpatient care is higher than the overall mean score for all providers, and the increase is only 4 dollars at the TIN level, which still demonstrates that providers will not be severely penalized for having a few high-cost events. The mean score for acute inpatient stay at the TIN level does not substantially increase until it is at Q3 or Q4, demonstrating that the measure differentiates performance based on the relative frequency of high-cost events. At the TIN-NPI level, acute inpatient stays at the Q2 frequency also had a higher mean score than the overall mean score for all providers.

---

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

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

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

Even though the Q2 mean score is substantially higher than the overall mean score, the mean score still follows the trend of increasing as the frequency of acute inpatient stays increases. This also demonstrates that the measure will take into account the frequency of high-cost events compared to peers to differentiate performance of providers, rather than the simple presence of high-cost events during the performance period.

Table 4: Mean and Standard Deviation of Score Stratified by the Frequency of Observing High-Cost Events

High-Cost Events	TIN			TIN-NPI		
	Quartiles (Frequency Range)	Mean Score	Score Standard Deviation	Quartiles (Frequency Range)	Mean Score	Score Standard Deviation
All Episodes	N/A	$1,451	$533	N/A	$1,441	$542
Emergency Department	0%	$1,315	$595	0%	$1,263	$526
	Q1 (0.3% - 4.8%)	$1,379	$559	Q1: (0.4% - 4.8%)	$1,355	$535
	Q2: (4.8% - 8.2%)	$1,417	$494	Q2: (4.8% - 8.3%)	$1,420	$512
	Q3: (8.2% - 12.1%)	$1,481	$487	Q3: (8.3% - 12.5%)	$1,483	$513
	Q4: (12.1% - 50%)	$1,588	$526	Q4: (12.5% - 56.5%)	$1,626	$558
Acute Inpatient Stay	0%	$1,433	$550	0%	$1,416	$536
	Q1: (0.1% - 0.7%)	$1,363	$316	Q1: (0.2% - 1.8%)	$1,453	$462
	Q2: (0.7% - 1.4%)	$1,455	$389	Q2: (1.8% - 2.9%)	$1,563	$516
	Q3: (1.4% - 2.6%)	$1,534	$494	Q3: (2.9% - 4.0%)	$1,670	$555
	Q4: (2.6% - 27.6%)	$1,667	$594	Q4: (4.0% - 21.9%)	$1,736	$621

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

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

In general, the direction of correlations indicate the following:

- Positive correlations with quality measures indicate that clinicians providing better quality care on that particular metric tend to also have lower costs. That is, clinicians who have high rates of performing specific quality actions (as measured through process measures) or achieve better patient health outcomes (as measured through outcomes

measures) tend to have lower costs of care. As such, these associations could represent ways to lower costs while also providing high-quality care
- A negative correlation between a cost and quality measure does not indicate an absence of cost improvement potential consistent with high-value care. This is because other approaches to improving cost performance (e.g., patient education) may not be captured by the selected quality measure.

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

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

There are very low numbers of TINs and TIN-NPIs that have both the quality and cost measure, providing a limited sample for this analysis. There are 17,237 TINs and 23,927 TIN-NPIs that meet the testing volume threshold for Major Depressive Disorder; across all the quality measure pairs tested, the highest number of overlapping providers was 1,652. There are also differences in patient cohort where the quality measures listed below apply to patients over the age of either 12 or 18; in contrast, the cost measure uses Medicare claims data so mainly covers patients over the age of 65. Q134 also excludes patients with a diagnosis of major depressive disorder as it is intended to assess screening for patients before they are diagnosed. In general, the correlations are very small in either direction, suggesting that cost performance can be improved without negatively impacting quality. There are no strong inverse correlations, suggesting that care stinting is not a concern within the limitations of this analysis.

Table 5. Correlation between Quality Measures and Cost Measure

Related Quality Measure (Type)	TIN or TIN-NPI	Number of Entities with Both Cost and Quality Measures Available	Pearson Correlation	P-value
Q009 Anti-Depressant Medication Management (Process)	TIN	108	0.094	0.336
	TIN-NPI	140	0.165	0.051
Q107 Adult Major Depressive Disorder (MDD): Suicide Risk Assessment (Process)	TIN	149	-0.073	0.376
	TIN-NPI	292	0.092	0.117
Q134 Preventive Care and Screening: Screening for Depression and Follow-Up Plan (Process)	TIN	775	-0.051	0.153
	TIN-NPI	1,652	-0.003	0.900
Q370: Depression Remission at Twelve Months (Outcome)	TIN	201	-0.139	0.050
	TIN-NPI	347	-0.095	0.077

2.5 Performance Gap

Table 6 shows the distribution of the measure score for clinicians and clinician groups. These results align with expectations based on our review of the literature and demonstrate that there is a performance gap in cost measure performance at both the clinician and clinician group levels. The Major Depressive Disorder cost 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. The variation in the measure score, indicated by the interquartile range and standard deviation, is in the hundreds of dollars. The results suggest that there is opportunity for improvement in performance across providers.

Table 6. Distribution of the Measure Score

Metric	TIN	TIN-NPI
Mean Score	$1,451	$1,441
Score Interquartile Range (IQR)	$561	$605
Standard Deviation	$533	$542
Coefficient of Variation	0.4	0.4
Score Percentile		
10th	$916	$880
25th	$1,111	$1,082
50th	$1,352	$1,347
75th	$1,672	$1,687
90th	$2,091	$2,100

2.6 Risk Adjustment and Stratification

2.6.1 Discrimination

Discrimination is a statistical criterion that evaluates the measure's ability to distinguish high-cost episodes from low-cost episodes, or the ability to explain the variance in cost of individual episodes. The amount of variance explained is estimated by the R-squared metric with the range between 0 and 1. The R-square for the measure is 0.12, and 0.12 after adjusting for the model's complexity based on the number of risk adjustors used. In other words, 12.4% of the variation in the actual observed cost of episodes is explained by the risk adjustment model and sub-group stratification.

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

2.6.2 Calibration

Calibration evaluates the consistency of the measure in estimating episode cost across the full range of resource use patterns in the population. Calibration is estimated by the average predictive ratios across groups within the population, specifically groups are partitioned by

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

Table 7 shows that the model has consistent predictive ratios across risk score deciles, with each decile having a predictive ratio between 0.98 and 1.01. The average predictive ratio for all risk deciles is 1.00, which demonstrates that the risk adjustment does not under- or over- predict across the full range of resource use patterns in the population.

Table 7. Predictive Ratio by Decile of Predicted Episode Cost

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

2.7 Social Risk Factor Analysis

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

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

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

1. Dual Medicare-Medicaid eligibility status:
    - Base model
    - Medicare-Medicaid eligibility status: full, partial, or non-dual status
2. All Socioeconomic Status (SES) Variables Model:

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

- Base model
- Dual Medicare-Medicaid eligibility status: full, partial, or non-dual status
- Sex: female or male
- Race: Asian, Black, Hispanic, North American Native, White, and other
- Neighborhood Income Level: low income (median income < 33rd percentile nationally), medium income (median income between 33rd and 66th percentiles), and high income (median income > 66th percentile)
- Neighborhood Education Rate: majority less than high school, high school, or greater than high school
- Neighborhood Employment Rate: less than or greater than 10%

3. Agency for Healthcare Research and Quality SES Index Model:
   - Base model
   - Dual Medicare-Medicaid eligibility status
   - Sex: female or male
   - Race: Asian, Black, Hispanic, North American Native, White, and other
   - AHRQ SES Index Score: calculated using the AHRQ's scoring algorithm, based on data from the American Community Survey

Table 8 shows the percent of providers who would see their performance shifted, based on their percentile ranking, using the new SRF models. There are marginal changes to the performance at both TIN and TIN-NPI reporting levels compared to the base model. Across the 3 models, over 80% TINs and TIN-NPIs would experience small and likely inconsequential swing in their performance, of less than 5% compared to the base model. Additionally, the performance ranking has Spearman correlation values of greater than 0.98 between the base model and each of the SRF models, which also suggests that adjusting for SRF variables does not fundamentally change how the measure evaluates provider performance.

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

Model	TIN or TIN-NPI	Proportion of Clinicians Affected at Various Levels of Performance Shift						
		-10% or more	-9% to -6%	-5% to -2%	-1% to 1%	2% to 5%	6% to 9%	10% or more
Model 1	TIN	0.3%	1.4%	10.9%	76.8%	10.6%	*	*
	TIN-NPI	0.2%	1.2%	10.5%	79.4%	8.7%	*	*
Model 2	TIN	0.9%	4.8%	22.7%	44.5%	21.0%	4.6%	1.6%
	TIN-NPI	0.8%	4.9%	23.5%	43.1%	21.5%	4.9%	1.4%
Model 3	TIN	2.6%	6.9%	21.8%	37.9%	21.5%	6.2%	3.0%
	TIN-NPI	2.4%	6.8%	22.7%	37.3%	21.7%	6.7%	2.6%

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

2.8 Impact of Exclusions

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

Overall, exclusion criteria decrease the distribution of observed cost of all episodes meeting trigger logic, from the mean of $2,019 to $1,428 at the TIN-level and $1,443 at the TIN-NPI level (Table 9). All of the exclusion criteria have higher mean observed cost than 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 $3,348 (Table 9), 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 $5,203 compared to $2,019 for all episodes meeting triggering logic (Table 9). The wide variability of observed episode costs for outlier cases also supports their exclusion. At the 10th percentile the outlier cases observed cost is $157 and at the 90th percentile the observed cost is $10,049.

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

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 trigger logic (Table 9), which suggests that they may have distinct resource use patterns from a typical major depression episode.

Table 9: Cost Statistics for Measure Exclusions

Exclusion Criteria	Episodes		Observed Episode Cost					
				Percentile				
	Count	Percent of All Episodes Meeting Trigger Logic	Mean	10th	25th	50th	75th	90th
All Episodes Meeting Triggering Logic	2,787,664	100.0%	$2,019	$219	$394	$859	$2,068	$4,857
Episode Length Less Than 1 Year	159,118	5.7%	$3,929	$543	$959	$1,954	$4,158	$8,147
Beneficiary Death in Episode	282,701	10.1%	$3,348	$426	$780	$1,634	$3,557	$7,090
Outlier Cases	43,876	1.6%	$5,203	$157	$353	$6,662	$10,049	$10,049
No Attributed Clinician (TIN-NPI Reporting Only)	148,467	5.3%	$2,692	$361	$642	$1,377	$3,063	$6,293
Presence of Bipolar Disorder Pre-Trigger	136,422	4.9%	$4,039	$380	$778	$1,747	$4,374	$10,354
Presence of Bipolar Disorder Post-Trigger	222,314	8.0%	$4,084	$386	$786	$1,772	$4,428	$10,279
Presence of Drug/Alcohol Psychosis Pre-Trigger	9,828	0.4%	$4,944	$338	$741	$1,773	$4,553	$12,050
Presence of Drug/Alcohol Psychosis Post-Trigger	17,027	0.6%	$6,159	$398	$875	$2,189	$6,465	$16,555
Presence of Schizophrenia Pre-Trigger	86,460	3.1%	$5,148	$465	$942	$2,092	$5,557	$13,867
Presence of Schizophrenia Post-Trigger	127,775	4.6%	$5,063	$467	$941	$2,100	$5,470	$13,265
TIN does not Meet Testing Volume Threshold	457,209	16.4%	$2,292	$212	$406	$1,027	$2,681	$5,597
TIN-NPI does not Meet Testing Volume Threshold	1,446,589	51.9%	$2,030	$207	$363	$819	$2,208	$4,931
Reportable Episodes (if all clinicians reported as TIN at the Testing Volume Threshold)	**1,791,844**	**64.3%**	**$1,428**	**$208**	**$349**	**$687**	**$1,652**	**$3,571**
Reportable Episodes (if all clinicians reported as TIN-NPI at the Testing Volume Threshold)	**913,553**	**32.8%**	**$1,443**	**$217**	**$370**	**$715**	**$1,655**	**$3,577**