NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

Purple text represents the responses from measure developers. Red text denotes developer information has changed since the last measure evaluation review. Some content in the document is from Measure Developers.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3574

De.2. Measure Title: Medicare Spending Per Beneficiary (MSPB) Clinician

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services

De.3. Brief Description of Measure: : The MSPB Clinician measure assesses the cost to Medicare for services by a clinician and other healthcare providers during an MSPB episode, which focuses on a patient's inpatient hospitalization. The MSPB episode spans from 3 days prior to the hospital stay ("index admission") through to 30 days following discharge from that hospital. The measure includes the costs of all services during the episode window, except for a limited list of services identified as being unlikely to be influenced by the clinician's care decisions and that are considered clinically unrelated to the management of care. The episode is attributed to the clinician(s) responsible for managing the beneficiary's care during the inpatient hospitalization. The MSPB Clinician measure score is a clinician's average risk-adjusted cost across all episodes attributed to the clinician. The beneficiary populations eligible for the MSPB Clinician measure include Medicare beneficiaries enrolled in Medicare Parts A and B during the performance period.

IM.1.1. Developer Rationale: : Increases in Medicare spending have been an important driver of rising total health care expenditures in the United States. [1] Given that the inpatient hospital setting is a significant contributor to overall Medicare spending, gauging the efficacy of this spending requires measuring the cost performance of clinicians providing care at hospitals. [2] The MSPB Clinician measure provides valuable context for such progress by measuring costs of care from a holistic perspective at the beneficiary level and offering a tool to control rising health care costs. [3]

[1] "National Health Expenditure Projections, 2017-2026." US Centers for Medicare & Medicaid Services, 2018.

[2] "Report to the Congress: Medicare Payment Policy." MedPAC, 2018. http://www.medpac.gov/docs/default-source/reports/mar18_medpac_entirereport_sec.pdf.

[3] Data Book: Health Care Spending and the Medicare Program." MedPAC, 2017. http://www.medpac.gov/-documents-/data-book.

De.1. Measure Type: Cost/Resource Use

S.5. Data Source: Asssessment Data

Claims

Enrollment Data

Other

S.3. Level of Analysis: Clinician : Group/Practice, Clinician : Individual

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. High impact or high resource use:

The measure focus addresses:

– a demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality).

AND

1b. Opportunity for Improvement**:**

Demonstration of resource use or cost problems and opportunity for improvement, i.e., data demonstrating considerable variation cost or resource across providers

---

1a. High Impact or high resource use.

- This measure calculates the average risk-adjusted cost to Medicare for services by a clinician and other healthcare providers during an Medicare Spending Per Beneficiary (MSPB) episode, which focuses on a patient's inpatient hospitalization. The measure is specified at the individual clinician and clinician group level by calculating the clinician's average risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician.
- The MSPB episode spans from 3 days prior to the hospital stay ("index admission") through to 30 days following discharge from that hospital
- The developer points out that inpatient hospital setting is a significant contributor to overall Medicare spending. Specifically, the developer cites a 2017 MedPAC report indicating that inpatient hospital spending accounted for 22 percent of total Medicare spending in 2015 and represented the second largest Medicare spending category in 2015.

1b. Opportunity for Improvement.

- The developer provides data demonstrating that MPSB episodes have a range of cost performance at the TIN and the TIN NPI level. Specifically, the interquartile range of performance for TIN level scores is $2,049 and mean performance of $19,194 for 19,213 group practices. The interquartile range of performance for TIN-NPI is $2,335, and mean performance of $19,741 for 126,628 practitioners.

Questions for the Committee:

- *Has the developer demonstrated this is high impact, high-resource use area to measure?*
- *Is there a sufficient variation in performance across hospitals that warrants a national performance measure?*

Staff preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low**

☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b)

1a. High Impact or High Resource Use: Has the developer adequately demonstrated that the measure focus addresses a high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality)?

Comments:

- **yes**
- **No**
- **Yes.**
- **this measure touches nearly all Medicare spending that starts with a hospital admission and 30 days post. Yes, affects large numbers.**
- **Yes.**
- **Yes.**
- **developer points to 2017 MedPAC report (which is based on 2015 data) that found inpatient hospital spending accounted for 22% of total Medicare spending and was the second largest Medicare spending category in 2015. There's no question that inpatient spending is a factor in high resource use, but I would like to see more up-to-date data from the developer to demonstrate measure addresses the issue.**
- **No concerns**
- **Yes, high resource use**

1b. Opportunity for improvement: Was current performance data on the measure provided? Has the developer demonstrated there is a resource use or cost problem and opportunity for improvement, i.e., data demonstrating, considerable variation in cost or resource use across providers?

Comments:

- **yes, if one accepts risk adjustment model**
- **No**
- **Yes. The interquartile range of performance for TIN level scores is $2,049 and mean performance of $19,194 for 19,213 group practices. The interquartile range of performance for TIN-NPI is $2,335, and mean performance of $19,741 for 126,628 practitioners. These variations in performance as reflected in IQRs indicate room for improvement.**
- **Yes, there is variation in spending that measure developer demonstrates.**
- **Yes.**
- **Yes, the developer demonstrated variation in cost and resource use across providers.**
- **TIN Level (Group) scores at 10th percentile $17,152 and 90th percentile $21,385 – not a huge variation; similar for TIN/NPI Level (Clinician) scores – at 10th percentile $17,504 and 90th percentile $22,067. Considering the measure is focused on spend from 3 days prior to the index admission through 30 days post discharge (including the inpatient stay) this seems low for variation.**

- **No concerns**
- **Yes**

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing

2b. Validity: Alignment of Specifications with Intent (includes threats to validity [e.g., Attribution**,** costing method**,** missing data**]) Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Multiple Data Sources; and Disparities.**

Measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators: Bijan Borah, MSc, PhD, Jack Needleman, PhD, Jennifer Perloff, PhD, Zhenqiu Lin, PhD, Jeffrey Geppert, EdM, JD, Eugene Nuccio, PhD, Christie Teigland, PhD, Susan White, PhD, RHIA, CHDA, Ronald Walters, MD, MBA, MHA, MS **(**Evaluation A: Methods Panel**)**

Methods Panel Individual Reliability Ratings: H-1; M-4; L-3; I-0 (Pass)
Methods Panel Individual Validity Ratings: H-0; M-5; L-3: I-0 (Pass)

- The developer provided responses to the concerns raised by the SMP, which can be found in the SMP Spring 2020 Discussion Guide on page 89 – 90.

Measure evaluated by Technical Expert Panel? ☐ **Yes** ☒ **No**

Reliability

2a1. Specifications:

The measure is well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability. All measures that use the ICD classification system must use ICD-10-CM.

2a2. Reliability testing:
Demonstration that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

2a2. Reliability Testing:

- The developer conducted signal-to-noise analysis and split sample reliability testing.
- Reliability scores were a mean of 0.78 and standard deviation of 0.13 for 19,213 TIN's and a mean of 0.70 with a standard deviation of 0.11 for 126,628 TIN-NPI's.
- Split sample intraclass correlation coefficients were 0.66 for TIN and 0.60 for TIN-NPI.
- Increasing size of practice groups was correlated with an increased reliability score. When examined by clinician group size, the average reliability score ranged from 0.70 (1 clinician) to 0.90 (21+ clinicians). The ICC for the overall sample was 0.76 with 95% confidence interval of 0.75-0.77.

Questions for the Committee regarding reliability:

- *Do you have any concerns with the reliability testing that was not identified by the Scientific Methods Panel?*
- *Would the Committee like to accept the SMP vote on reliability?*

Guidance from reliability algorithm

(Box 1) Are specifications precise, unambiguous, and complete? YES → (Box 2) Was empirical testing conducted using statistical tests with the measure as specified? YES → (Box 4) Was reliability testing at the score level? YES → (Box 5) Was the method appropriate? YES → (Box 6) Moderate certainty of measure reliability → MODERATE

Staff Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 2a: Reliability

2a1. Reliability – Specifications: Describe any additional concerns you have with the reliability of the specifications that were not raised by the Scientific Methods Panel:Describe any data elements that are not clearly defined:Describe any missing codes or descriptors:Describe any elements of the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) that are not clear:Describe any concerns you have about the likelihood that this measure can be consistently implemented:

Comments:

- **N/A**
- **The terms in Step 6 on page 11 need to be more directly linked to the discussion in Step 5, e.g. the "risk-adjusted episode cost ratio" need to be directly defined. Time periods used seem arbitrary and should be condition-specific. Death is endonenous here as poor care results in more death. Analysis should include all patients, and measure should be reported in conjuncture with provider-specific mortality rates across the episodes attributed to a provider. This will enable decision-makers to assess the tradeoffs. Also, "transfer-outs" is a decision of the care provided at the initial hospital and should be included in the measure. "Transfer-ins" are a distinct discussion and seem to be sensibly excluded here.**
- **No additional concerns other than those raised by the Scientific Methods Panel.**
- **Concern about this measure at the level of the individual physician and a few high cost cases leading to instability in the measure at that level. Article by Mehrotra et al. finds measuring cost at clinician level to not be all that reliable. We see this in the reliability stats provided by measure developer. I was not clear (when looking at the excel sheet provided by measure developer) that shows the variables entering the risk adjustment model and the coefficients. It looks like transplants are included. Is that true? I would think these cases are very different from everything else and should be set aside.**
- **No concerns**
- **No concerns that were not raised by the Scientific Methods Panel.**
- **none**
- **No concerns**
- **No concerns**

2a2. Reliability – Testing: Has the developer demonstrated that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers?Describe any additional concerns you have with the reliability testing results or approach that were not raised by the Scientific Methods Panel.

Comments:

- **Share concerns with SMP that S/N and split sample statistics are too low to make assessment at individual level of performance for small practices. Average reliability may be okay, but process seems to have been to pick sample sizes for inclusion before examing reliability rather than making sample sizes consistent with high enough S/N scores.**
- **Yes**

- **None other than those raised by the SMP.**
- **agree with SMP that reliability isn't strong here. A lot of random variation entering into the calculation, especially at clinician level. regarding reliability testing, this statement is very problematic: reporting case minimum have reliability greater than or equal to 0.4, the standard that CMS generally considers as the threshold for 'moderate' reliability. This is incorrect. Anything lower than 0.6 reliability is considered poor and reflects more noise than signal. I have concerns that 50% of all NPIs are below a reliability threshold of 0.69 (between 0.6 and 0.7 is considered low reliability). At a threshold of 35 episodes, this measure doesn't meet reliability for half the physicians being measured, and is weak in discriminating performance. The pearson coefficient for split sample is on the weaker side (of around 0.66) showing disagreement a good share of the time in test/retest of classifying providers.**
- **No concerns.**
- **No concerns that were not raised by the Scientific Methods Panel.**
- **Moderate; would like more clarification on whether reliability was actually tested for clinician individuals (it seems as though the SMP panel wasn't clear on that and it appeared only to reliability test for groups) There is a question of whether the results on reliability for MSBP for clinicians are yield the right results. There are concerns with the lack of information on reliability results below the 25th percentile. CMS states that it considers 0.4 to be the threshold for moderate reliability and that 100% of practices with at least 35 episodes meet it. The minimum acceptable threshold should be higher than 0.4. (some believe it should be 0.7).**
- **No concerns**
- **No concerns**

Validity

2b1. <u>Specifications align with measure intent</u>:

The measure specifications are consistent with the measure intent and captures the most inclusive target population.

2b2. <u>Validity Testing</u>:

Demonstration that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided.

2b3. <u>Exclusions:</u>

Exclusions are supported by the clinical evidence**,** AND/OR There is a rationale or analysis demonstrating that the measure results are sufficiently distorted due to the magnitude and/or frequency of then on-clinical exclusions; AND Measure specifications for scoring include computing exclusions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); AND If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

2b4. <u>Risk Adjustment:</u>

For resource use measures and other measures when indicated: an evidence-based risk-adjustment strategy is specified and is based on patient factors (including clinical and sociodemographic risk factors) that influence the measured outcome and are present at start of care, and has demonstrated adequate discrimination and calibration, OR rationale/data support no risk-adjustment/-stratification.

2b5. Meaningful Differences:

Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/ clinically meaningful differences in performance.

2b6. Multiple Data Sources:

If multiple data sources/methods are specified, there is demonstration that they produce comparable results.

2c. Disparities: If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender), OR rationale rationale/data justifies why stratification is not necessary or not feasible.

2b1. Specifications Align with Measure Intent:

- The measure focuses on costs associated with a patient's inpatient hospitalization period, in which a given episode period is indicated as three days before hospitalization to 30 days after discharge.
- The SMP raised the concern about attribution to multiple clinicians and whether a care episode could be attributed to multiple clinician groups and multiple clinicians.
- The SMP raised a concern that during these surgical DRGs, the clinician(s) caring for the main disease process are the ones that are driving the care for the patient as opposed to the proceduralist who performed the primary procedure. For example, in an episode where a patient who has a prolonged intubation during an ICU stay and requires a tracheostomy, it would be unfair to attribute the proceduralist who performs the tracheostomy as they may only see the patient for this single procedure.

2b2. Validity Testing:

- **Face Validity Testing**
    - Both face validity and empirical validity testing were conducted.
    - Face validity comprised of administering a structured process for gathering detailed input from recognized clinician experts on inpatient care. Multiple expert panels informed the face validity of the measure at different time points: a technical expert panel (TEP), the MSPB service refinement group, and stakeholder feedback from national field testing.
    - The developer states that "Out of 15 respondents to the survey, 14 (93%) agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost effectiveness."
- **Empirical Testing**
    - For establishing empirical validity, the measure developer sought to confirm the expectation that the MSPB Clinician measure captures variation in service utilization by examining differences in risk-adjusted cost for known indicators of resource or service utilization (e.g., inpatient readmissions or post-acute care) through the ratio of observed-to-expected cost ("O/E cost ratio").
    - Results of empirical validity testing found the mean O/E cost ratio for episodes with downstream acute readmission is 1.58, compared with 0.91 for episodes without downstream acute readmission indicating the measure is able to accurately capture higher resource use related to readmission. The mean O/E cost ratio for episodes with post-acute care (PAC) is 1.20, while for episodes without PAC is 0.80 as hypothesized
    - Additionally, the developer tested whether the MSPB measure is capturing variation in provider cost in the manner intended by evaluating how different types of cost arising from homogenous clinical themes (e.g., acute inpatient services, inpatient readmission, post-acute care etc.) impact risk-adjusted measure scores (Table 5).

Table 5. Pearson Correlation Statistics between Costs for Clinical Themes with Risk-Adjusted and Expected Costs

Clinical Theme	Average Cost of Grouped Clinical Theme	Pearson Correlation With Risk Adjusted Cost	Pearson Correlation With Predicted Cost
Acute Inpatient Services: Index Admission*	$11,561	0.08	0.87
Acute Inpatient Services: Readmission	$8,863	0.47	0.04
Emergency Services Not Included in Hospital Admission	$739	0.08	-0.01
Outpatient E&M Services, Procedures, and Therapy	$850	0.26	0.01
Post-Acute Care: Home Health	$1,933	-0.18	0.01
Post-Acute Care: IRF/LTCH	$22,518	0.15	0.55
Post-Acute Care: SNF	$11,181	0.34	0.06

- Some SMP members questioned the strength of the correlations, noting that the correlation between predicted value and six different "clinical themes (e.g., PAC settings) was low (< 0.10) in all cases except PAC IRF/LTCH." And that the correlation with risk adjusted value and six different "clinical themes was also not high—and was negative (-0.18) with PAC Home Health."

2b3. Clinical Inclusions and Exclusions/Evidence to Support Clinical Logic

- The developer conducted an exclusion analysis on the exclusion criteria, finding that of the 10,658,462 episodes, 59.1% remained and had a lower observed cost.
- Some SMP members raised concerns with the outlier exclusions (i.e., episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution) – specifically, should episodes with very low or very high residuals be excluded?

2b4/2c. Risk adjustment

- The developer used a statistical risk model with 109 risk factors and performed stratification by 26 risk categories.
- The developer states that the risk adjustment model for the MSPB Clinician measure "broadly follows the CMS-HCC risk adjustment methodology used in the Medicare Advantage (MA) program." Severity of illness is measured using Hierarchical Condition Category (HCCs), indicators of enrollment and long-term care status, and disease interactions. There also 12 categorical age variables included in the model.
- The range of R-squared values for the MSPB clinician cost measure risk adjusted models ranges between 0.09 – 0.64 across the MDCs. The adjusted R-squared range is 0.09 – 0.63Windorization – Standard methodology in cost measurement.

2b5: Meaningful Differences

- The developer reports that the variability of clinician scores (Table 7), with 20% significantly lower than the national mean and 12.4% higher by TIN. For TIN-NPI the numbers are 15.3% and 8.0%, respectively.

Table 7. Proportion of Measure Scores Statistically Significantly Different From the National Average

Provider Level	# of Providers	Statistically significantly lower than national mean		Not statistically significantly different from national mean		Statistically significantly higher than national mean	
		#	%	#	%	#	%
TIN	19,213	3,835	20.0%	12,996	67.6%	2,382	12.4%

TIN-NPI	126,628	19,326	15.3%	97,138	76.7%	10,164	8.0%

Questions for the Committee regarding validity:

- *Did the developer's submission adequately demonstrate empirical validity?*
- *Are the exclusions appropriate?*
- *Are the strengths and directions of the correlations acceptable?*
- *Are there any concerns with the developer's approach to determining social risk factors for inclusion in the risk model?*

Guidance from validity algorithm:

(Box 1) Were threats to validity addressed? YES → (Box 2) Was empirical testing conducted using statistical tests with the measure as specified? YES → (Box 5) Was validity testing at the score level? YES → (Box 6) Was the method appropriate? YES → (Box 7) Moderate certainty of measure validity → MODERATE

Staff preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 2b: Validity

2b1. Validity –Testing: Describe any concerns you have with the testing approach, results and/or the Scientific Methods Panel and NQF-convened Clinical Technical Expert Panel's evaluation of validity:Describe any concerns you have with the consistency of the measure specifications with the measure intent:Describe any concerns regarding the inclusiveness of the target population:Describe any concerns you have with the validity testing results:Does the testing adequately demonstrate that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided?

Comments:

- **Lack of detail in discussion of methods for identifying Unrelated Services for exclusion.**
- **Validity should be more based on the actual concept needed to be measured for a given study. Comparing costs to healthcare utilization seem tautological. What else could be making "cost" move?**
- **None.**
- **Measure developer does not ultimately include social risk factors; however, when I look at the coefficients, the effects on spending look significant and as large as many other factors in the model. I do think that accountability at the clinician or small group level on this measure may lead to undesired effects of clinicians avoiding patients with social risk factors. There is a lot of variation across practices (TINs) and individual physicians in the extent to which they care for people with social risk factors. While it doesn't have much effect on average, it is at the tails of the distribution where the effects are more keenly felt. So in most cases risk adjustment doesn't change the scores for most providers (here or in other models for other measures), it does have impact for those with large fractions of patients with social risk factors. Another issue is that a key driver of the variation in this measure is SNF spending, and the question is whether the provider who is accountable for care in the inpatient setting is also accountable for spending in the SNF? This attribution logic is questionable.**
- **No concerns.**
- **No concerns not raised by the Scientific Methods Panel.**
- **agree with concern about attribution to multiple clinicians/clinician groups – as attribution for medical DRGs based on 30% of the billed services during the episode – this developer says this acknowledges team-based care but it seems challenging for using the measure to improve performance if those clinicians/clinician groups are not aware of the coordination efforts to improve (if it is one-off shared patients and not routine "sharing" of responsibility for patient costs)**
- **No concerns**
- **No concerns**

2b5a. Threats to Validity: Meaningful Differences: Describe any concerns with the analyses demonstrating meaningful differences among accountable units:

Comments:

- **If I believed the underlying logic of including all services in the post 30 day period, and that the risk adjustment model accurately adjusted for appropriateness of SNF services, then I would say yes, the differences are sufficiently large to be meaningful, but I have questions about both, and therefore don't have confidence in the differences across entitities or O/E that are estimated.**

- **None.**
- **None.**
- **yes, there are differences, largely driven by post-acute care spending. Variation is in SNF spending, which seems to be a far reach for the clinician assigned the episode. agree that showing that the measure is correlated with its component parts is a weaker test of validity.**
- **No concerns.**
- **No concerns not raised by the Scientific Methods Panel.**
- **There seems to be a carve out for certain procedure DRGs to attribute based on medical methodology; should those procedures simply be a carve out as a whole rather than using the medical methodology?**
- **No concerns**
- **No concerns**

2b5b. Threats to Validity: Missing Data/Carve-outs: Describe any concerns you have with missing data that constitute a threat to the validity of this measure:Carve Outs: Has the developer adequately addressed how carve outs in the data source are handled (or should be handled for other users)? For example, if pharmacy data is carved out (missing) from the data set, can a measure that focuses on cost of care the target clinical opulation still be valid?

Comments:

- **okay**
- **None.**
- **I wonder about the inclusion (if I'm understanding it correctly) of transplant services.**
- **No concerns.**
- **None.**
- **): Attribution: concern that attribution to multiple clinicians/clinician groups, especially on a retrospective basis, leads little information for how a clinician/clinician group could better coordinate to improve efficient use of resources/costs. Costing approach: Would want more detail about how hospital drug and innovative new technology care/costs come into the calculations; examples of things like CAR-T and impact. Challenge is how much inpatient care costs and post-discharge costs are in the control of a non-hospitalist clinician/clinician group.**
- **No concerns**
- **No concerns**

2b2. Additional threats to validity: attribution, the costing approach, or truncation: Describe any concerns of threats to validity related to attribution, the costing approach, or truncation (approach to outliers):Attribution: Does the accountable entity have reasonable control over the costs/resources measured? Is this approach aspirational (intending to drive change) or was it developed based on current state?Costing Approach: Do the cost categories selected align with the measure intent, target population and care settings? Is the approach for assigning dollars to resources agreeable?Truncation (approach to outliers): What is the threshold for outliers (i.e., extremely high cost or low cost cases) and are they handled appropriately?

Comments:

- **There are substantial issues raised in the SMP report about appropriate attribution. There is also the issue that this measures specs appear also to be used in a hospital level measure, so the same variance is being attributed to hospitals and physicians. I can accept that but would like discussion of the logic of dual attribution by the committee.**
- **Once again, algorithm parameters with regard to episode window should be condition specific. Atribution should be based on providers who help make key care decisions.**
- **None.**
- **Social risk factor exclusion seems problematic. What is unclear is did the measure developer is adjusting for within clinician disparity or just including the overall effects (say of dual status). The idea is not to wash away differences between providers, but to control for the within provider effects of social risk factors. The documentation was unclear on this front (or I wasn't successful in finding it in all the materials)**
- **Not a concern since the SMP reviewed the measures, but two questions since I am less familiar with these methodologies. In winsorization of predicted costs, why was the decision made to bottom-code only and not top-code? Why were outliers defined based on the residuals?**
- **Hospitals are being held accountable for the same index admissions and post-discharge periods as clinicians. The measure does not address the hospital's influence over this measure and the impact of a specific hospital's resources and the actions of the hospital staff. It is also unclear to what degree a clinician has control over certain aspects of care within a hospital.**
- **none**
- **No concerns**
- **No concersn**

2b3. Additional Threats to Validity: Exclusions: Describe any concerns with the consistency exclusions with the measure intent and target population:Describe any concerns with inappropriate exclusion of any patients or patient groups:

Comments:

- **Part C exclusion inherently leads to large proportion of excluded cases.**
- **None.**

- **I am concerned about the absense of adjustments for social risk factors. Also, I have concerns about the measure construction with creating the NPI/TIN's average cost ratio and then multiplying this against the national average cost ratio (which includes episodes that are dissimilar to the NPIs/TINs). Seems the measure should be better tailored to the specific types of episodes the TIN is treating when doing the multiplier to some national average. if I understand the measure logic, all clinicians nationally are the peer group, rather than those that treat similar types of episodes. That seems problematic.**
- **No concerns.**
- **None.**
- **No concerns**
- **No concerns**

2b4. Additional Threats to Validity: Risk Adjustment: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factors that were available and analyzed align with the conceptual description provided?Has the developer adequately described their rationale for adjusting or stratifying for social risk factors?Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Describe any concerns with the appropriateness of risk adjustment (case-mix adjustment) development and testing:Do analyses indicate acceptable results?

Comments:

- **I don't like the performance of the risk adjustment model re SNF care. The intent of this measure is attribute costs to clinicians providing inpatient care the costs of the inpatient care and post-acute care that flow from the hospitalization. Higher O/E costs are taken as an indicator of either excessive services relative to patient need or treatment of complications or other adverse consequences that flow from the care provided by the clinician in the hospital. This requires that the risk adjuster adequately control for higher/lower resource use due to patient condtion and comorbidities, and to appropriate prescription of post-acute services such as SNF, HH, and LTCH. It is not clear that the HCC's and other elements of the risk adjuster robustly differentiate appropriate from inappropriate use of expensive post-acute services. The risk adjustment model is extensive, and has many good characteristics: separate model for each MDC, inclusion of substantial number of disease/illness related indicators, inclusion of indicators for disability and prior institutionalization. However, I am concerned that it underestimates APPROPRIATE use of SNF. This concern is based on data reported in Table 5 Pearson Correlation Statistics between Costs for Clinical Themes with Risk-Adjusted and Expected Costs on page 13 of the testing document. This table shows a 0.06 correlation of SNF care with predicted cost, and 0.34 correlation with risk adjusted costs. By contrast, the correlation of predicted cost for IRF/LTCH is 0.55 and with risk adjusted cost 0.15. Some patients post-hospitalization will require SNF care, even as some SNF care is unnecessary or excessive. The risk adjuster should therefore adjust out needed or appropriate SNF care. The 0.06 correlation of SNF and predicted cost, while not zero, seems low to me. If there are MDCs with higher expected SNF care and the correlation for those MDCs of predicted costs to SNF costs was higher, this concern might be mitigated.**
- **Average aberations from risk-adjusted costs should be reported, not average risk-adjusted costs.**
- **None except that I agree with the SMP member # 3 that SES characteristics should have been included in the final risk-adjustment model.**
- **social risk factor issue I flagged previously. Think adjusting for within clinician differences takes on greater importance at the individual clinician level**

- **The range of r-squared across the different MDCs seems broad and in some cases low. I know the developer provided rationale for the lower r-squared, but am still uncertain whether the risk adjustment is sufficient for the purposes of clinician accountability**
- **The developers describe a separate model for each MDC with adjusted R-squared range is 0.09 – 0.63. This would indicate there is a lot of variability in the model fit and potentially some cases where prediction may be too poor to use in a measure incorporated into a payment program.**
- **Broadly follows CMS HCC risk adjustment model which is not sensitive to social risk factors as that data is not typically available from claims, though the developer seems to report that social risk factors are included in the risk model. Developer seems to suggest that gender and dual status are sufficient SES proxies, which are included in HCC model, and did not choose to go further and use AHRQ SES index. There are also additional concerns in light of COVID-19 pandemic with HCCs as it is based on prior year claims. Will risk adjustment be really out sync due when 2020 is the year used for assigning HCCs? Note that there are restrictions on diagnoses for HCC scores with telehealth services in some cases, in addition to the reduction of services broadly and the predicted uptick of services in 2021 due to delayed care. ? It does not appear to have been addressed, just seems to be HCC model is enough and workable. The risk adjustment model is not adequate due to R squared results ranging from 0.09 to 0.64 across groupings. In addition, it is not adequately tested and adjusted for social risk factors. The developer tested social risk factors after adjusting for clinical risk factors rather than assessing the impact of both clinical and social risk factors in the model at the same time. This could impact how each variable performs in the model**
- **No concerns**
- **No concerns**

Combined Scientific Methods Panel Preliminary Analysis of Scientific Acceptability

Measure Number: 3574

Measure Title: Medicare Spending Per Beneficiary (MSPB) Clinician

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☒ **Cost/Resource Use**

☐ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☒ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**
☐☒ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☐ **Registry Data**
☐☒ **Enrollment Data** ☒☐ **Other:** Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment **(Panel Member #1) (MDS) (Panel Member #2)**

Level of Analysis:

☒☐ **Clinician: Group/Practice** ☒ **Clinician: Individual** ☐ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☐ **Population: Regional and State**
☐ **Integrated Delivery System** ☐ **Other**

Measure is:

☒ **New** ☐ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? ☐☒ Yes ☒☐ No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

   ***NOTE**: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   Panel Member #3 The MSPB Clinician measure assesses the cost to Medicare for services by a clinician and other healthcare providers during an MSPB episode, which focuses on a patient's inpatient hospitalization. The MSPB episode spans from 3 days prior to the hospital stay ("index admission") through to 30 days following discharge from that hospital. The measure includes the costs of all services during the episode window, except for a limited list of services identified as being unlikely to be influenced by the clinician's care decisions and that are considered clinically unrelated to the management of care. The episode is attributed to the clinician(s) responsible for managing the beneficiary's care during the inpatient hospitalization. The MSPB Clinician measure score is a clinician's average risk-adjusted cost across all episodes attributed to the clinician. The beneficiary populations eligible for the MSPB Clinician measure include Medicare beneficiaries enrolled in Medicare Parts A and B during the performance period.

   S.7.1 The MSPB Clinician measure is a clinician's average risk-adjusted cost across all episodes attributed to the clinician (TIN-NPI) or clinician group (TIN). The measure population is defined by admission to an inpatient hospital. The episode window starts 3 days prior to this index admission and ends 30 days after discharge. Episodes are attributed to clinicians and clinician groups based on the Part B services provided during a medical or surgical Medicare Severity Diagnosis-Related Group (MS-DRGs) inpatient admission. The costs of all services occurring during the episode window - except for a limited set of services that are not clinically related to the management of care for the episode - are summed to obtain each episode's standardized observed cost. A regression model is applied to the risk adjustment variables to estimate the expected cost of each episode.

   The cost measure is calculated as a the ratio of standardized observed cost to expected cost averaged across all of a clinician or clinician group's attributed episodes to obtain the average episode cost ratio. The average episode cost ratio is multiplied by the national average observed episode cost to generate a dollar figure for the cost measure score.

   S.7.2---STEP 1: Define and Trigger Episodes

   Episodes are opened, or triggered, by admissions to inpatient hospitals. The episode window starts 3 days prior to this index admission and ends 30 days after the hospital discharge. There is a 90-day lookback period before the episode start date. This period is used to check beneficiary enrollment information for episode exclusions and beneficiary pre-existing health characteristics used for risk adjustment.

   STEP 2: Attribute Episodes to Clinicians

   Attribution is the process of determining which clinician groups and clinicians are responsible for an episode. The MSPB Clinician measure utilizes two attribution methods for medical and surgical MS-DRG episodes:

   •Medical episodes (i.e., episodes for which the index admission has a medical MS-DRG) are attributed to any clinician/clinician group responsible for managing the medical condition during the inpatient stay. Specifically, the episode is attributed first to the TIN that bills at least 30 percent of E&M codes found on Part B Physician/Supplier claims during the inpatient stay. The episode is then attributed to the TIN-NPI who billed at least one E&M service that was used to determine the episode's attribution to the TIN.

   •Surgical episodes (i.e., episodes for which the index admission has a surgical MS-DRG) are attributed to the clinician/clinician group performing the main procedure during the inpatient stay. Specifically, the episode is attributed to the TIN and TIN-NPI who billed any related surgical procedure on Part B

Physician/Supplier claims during the inpatient stay. The full list of Current Procedural Terminology / Healthcare Common Procedure Coding System (CPT/HCPCS) codes determined as related to each surgical MS-DRG can be found in the Measure Codes List file. See Section S.1. for link to Measure Codes List.

A few select surgical MS-DRGs are attributed using the 30% E&M rule, rather than a main procedure. During these surgical DRGs, the clinician(s) caring for the main disease process are the ones that are driving the care for the patient as opposed to the proceduralist who performed the primary procedure. For example, in an episode where a patient who has a prolonged intubation during an ICU stay and requires a tracheostomy, it would be unfair to attribute the proceduralist who performs the tracheostomy as they may only see the patient for this single procedure.

STEP 3: Exclude Clinically Unrelated Services to Calculate Episode Observed Cost

All Medicare Part A and Part B concurrent to the episode window are considered for inclusion toward the episode, with exceptions for services that are unlikely to be influenced by the clinician's care decisions. Services unlikely to be influenced by the clinician's care decisions are excluded based on rules developed by the MSPB Service Refinement Workgroup (discussed in Section S.8.3). The service exclusion rules are defined specific to the Major Diagnostic Category (MDC) of the index admission. The service exclusion codes and logic for services deemed clinically unrelated can be found in the "SE_[General/Post]_[Service_Category]" tabs of the Measure Codes List file (see Measures Codes List linked in Section S.1). After applying service exclusions, the standardized Medicare allowed amounts for the services included in each episode are summed to obtain the standardized episode observed cost.

Payment standardization is a process used by CMS to adjust the allowed charge for services to facilitate comparisons of resource use by removing geographic differences (e.g., due to labor costs) and adjustments from special Medicare programs (e.g., graduate medical education and disproportionate share payments). By removing the effect of these factors, the payment standardization process preserves the differences in spending that are a result of healthcare delivery choices.

STEP 4: Exclude Episodes

A series of episode exclusions are applied to remove certain episodes from measure score calculation. Episodes are excluded from the MSPB Clinician measure if they meet any of the following conditions:

•Beneficiary has a primary payer other than Medicare during the episode window or in the 90-day lookback period

•Beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 90-day lookback period and episode window

•The beneficiary's death occurred during the episode.

•The index admission for the episode did not occur in either a subsection (d) hospital paid under the Inpatient Prospective Payment System (IPPS) or in an acute hospital in Maryland.

•The index admission for the episode is involved in an acute-to-acute hospital transfer (i.e., the admission ends in a hospital transfer or begins because of a hospital transfer).

•The index admission inpatient claim indicates a $0 actual payment or a $0 standardized payment.

After applying the exclusions outlined above, all remaining episodes are included in the calculation of the MSPB Clinician measure score.

Step 5: Calculate Expected Episode Costs Through Risk Adjustment

Risk adjustment is used to estimate episode expected costs in recognition of the different levels of care beneficiaries may require due to comorbidities, disability, age, and other risk factors. A separate risk adjustment model is estimated for episodes within each MDC, which is determined by the MS-DRG of the index admission. This model includes variables from the CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) 2016 Risk Adjustment Model and other standard risk adjustors to capture beneficiary characteristics.

Steps for defining risk adjustment variables and estimating the risk adjusted expected episode cost are as follows:

1) Define HCC and patient characteristic-related risk adjustors using Medicare Parts A and B claims in the 90-day lookback period from the episode start date.

2) Define other risk adjustors that rely upon Medicare beneficiary enrollment and assessment data as follows:

2a) Identify beneficiaries who are originally "Disabled without end-stage renal disease (ESRD)" or "Disabled with ESRD" using the original reason for joining Medicare field in the Medicare beneficiary enrollment database.

2b) Identify beneficiaries with ESRD if their enrollment indicates ESRD coverage, ESRD dialysis, or kidney transplant in the Medicare beneficiary enrollment database in the 90-day lookback period.

2c) Identify beneficiaries who are resident in a long-term care institution (90 days without having been discharged for 14 days) as of the episode start date using MDS assessment data.

3) Categorize beneficiaries into age ranges using their date of birth information in the Medicare beneficiary enrollment database.

4) Calculate an ordinary least squares (OLS) regression model to estimate the relationship between all the risk adjustment variables and the dependent variable, the standardized observed episode cost, to obtain the expected episode cost. A separate OLS regression is run for each episode MDC group nationally.

5) Winsorize the expected episode cost by assigning the value of expected episode cost at the 0.5th percentile of the distribution for episodes within the same MDC to all episodes with expected episode costs below the 0.5th percentile.

6) Renormalize values by multiplying each episode´s winsorized expected cost by the ratio of the MDC group´s average observed cost and the MDC group´s average winsorized expected cost.

7) Exclude episodes with outlier residuals to obtain finalized episodes with expected cost. This step is performed across all episodes regardless of the MDC group.

7a) Calculate each episode´s residual as the difference between the observed cost and the re-normalized, winsorized expected cost computed above.

7b) Exclude episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution.

7c) Renormalize all remaining episodes by multiplying their cost by the ratio of the average observed episode cost and the average winsorized expected cost when excluding outliers.

Step 6: Calculate Measure Scores

The MSPB Clinician measure is calculated for each clinician (TIN-NPI) or clinician group practice (TIN) by calculating the risk-adjusted episode cost ratio and multiplying the average cost ratio by the national average standardized episode cost. This method of cost ratio calculation allows for comparison of differences in observed and expected costs at the level of each individual episode before comparison at the clinician or clinician group level.

Specifically, the measure is calculated as follows:

•For each non-outlier episode, divide the episode's standardized observed cost by the episode's final expected cost to obtain the risk-adjusted episode cost ratio.

•Average the risk-adjusted cost ratios across all episodes for each TIN or TIN-NPI, and multiply this average cost ratio by the national average episode cost (all total standardized costs averaged over the universe of attributed, non-outlier episodes) to obtain the MSPB Clinician measure score for each TIN or TIN-NPI. Multiplying the ratio by the national average cost per episode is done to present the clinician's average cost measure score as a dollar amount rather than a ratio to be a more meaningful figure for clinicians.

Section S.8.1---- The measure aims to provide actionable information to clinicians providing care for beneficiaries during their hospital stay within the overall goal of enabling clinicians to provide cost-effective and high-quality care.

Section S.13.4-- From the MIPS CY 2020 performance period and onwards, the MSPB Clinician measure will be calculated and reported via confidential reports for TINs and TIN-NPIS with 35 or more episodes. Public reporting may be introduced for MIPS cost measures in the future.

2. **Briefly summarize any concerns about the measure specifications.**

Panel Member #2 This is a multi-step measure with many difficult to apply attribution, exclusion, and other factors, as well as a set of complex statistical models, etc. The measure could only be calculated with significant data and testing by CMS or similar entity with access to detailed programming specs and comprehensive data.

Panel Member #3

- Importance to Measure, Feasibility, and Usability and Use were blank
- Given the episode period (3 days before hospitalization to 30 days after discharge), how is this measure different from NQF #2158 MSPB Hospital cost measure and why does one measure attribute these costs to the clinician/clinician group vs. the hospital clinical staff?
- Measure "focuses on a patient's inpatient hospitalization." What about measuring the MSPB for a clinician's patient who DOES NOT go to the hospital? Some sort of stratification by clinician type (e.g., the number of patient inpatient hospitalizations would differ dramatically for a PCP vs. a gerontologist vs. an anesthesiologist vs. a surgeon).
- Care episode could be attributed to multiple clinician groups and multiple clinicians. What if a clinician has only 1 of 15 E&M codes and another clinician has 5 of the 15 E&M codes? The cost is computed based on the entire episode, not by E&M code. However, both clinicians would be assigned the same episode cost regardless of the number of (or relative expense of) the E&M codes.
- Attribution for surgical procedures seem to follow a different set of rules: A few select surgical MS-DRGs are attributed using the 30% E&M rule, rather than a main procedure. During these surgical DRGs, the clinician(s) caring for the main disease process are the ones that are driving the care for the patient as opposed to the proceduralist who performed the primary procedure. For example, in an episode where a patient who has a prolonged intubation during an ICU stay and requires a tracheostomy, it would be unfair to attribute the proceduralist who performs the tracheostomy as they may only see the patient for this single procedure.
- There is a lot winsorizing lower values, renormalizing, and removing outlier values (<1%/>99%). Will the reason why there are extremely high costs ever be investigated?
- If the CMS Comprehensive Error Rate Testing (CERT) program is effective—and given that there was no rationale offered for why Measure 3574 is important, why not use the CERT results and target clinicians and clinician groups that fall at the lower levels of this program? CMS also uses the Comprehensive Error Rate Testing (CERT) Program to ensure that Medicare payments are correct in accordance with coverage, coding, and billing rules. Between 2005 and 2017, CERT estimates that proper payment, which is payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments each year. The FY 2018 Medicare FFS program proper payment rate was 91.9 percent.[1] CMS continues to perform successful corrective actions and give providers additional education to ensure accurate billing.
- If the goal of the measure is to provide information about "cost-effectiveness and high-quality care", how is quality of care measured?
- Focus seems to be on costs associated with hospitalization period, but seem to neglect costs during the 30 days post discharge from the hospital. Would the measure push clinicians and clinician groups

to use home health agencies—due to their lower costs—than IRFs or SNFs for post-acute care? How will these PAC costs be captured?

- Sample size of 35 episodes may eliminate a large number of clinicians from the computation.

Panel Member #7 Understand that these are derived from lots of discussion about attribution but the inclusion/exclusion criteria and the attribution rules are still very confusing and subject to arbitrary assignments.

Panel Member #8 NA

Panel Member #9 My key concern is with the outlier exclusion (Step 6 – Contruction logic). Based on the measure specifications, episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution.

are excluded. Two factors make this approach particularly concerning: 1) This is a measure focusing on resource utilization, should episodes with very low or very high residuals be excluded? If the concern is with potentially undue influences of outliers, is exclusion the best available approach? 2) Winsorize predicted values: low predicted value below 0.5th percentile is already winsorized before calculation of residual. At minimum, the developer should report the distribution of outlier exclusion across providers to ensure that they don't concentrate in a limited number of providers

RELIABILITY: TESTING

Submission document: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☐☒ **Measure score** ☐ **Data element** ☐ **Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure**
   ☒ **Yes** ☐ **No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

   ☐ **Yes** ☐☒ **No** **X—Not applicable (Panel Member #3)**

   Panel Member #3 NOTE: Given the compound nature of the words prior to the comma—and the used of the "OR" conjunction, I really have no idea how to respond to this question. There is little to no likelihood that a Developer would submit a measure for review without attempting reliability testing of either the score (measure) or data elements. Whether the methods were appropriate or not cannot be answered until that information has been review (see item #6). Should item #5 be moved to the end of this section—and split apart into two questions? Or, should the item be simply eliminated as if the reliability testing was not done, then the responses to the questions that follow would lead to a "failure" or if the reliability testing methodology was not appropriate, then the measure would also "fail."

   While reliability is a necessary prerequisite for validity, demonstrating validity without any evidence that a measure or the data used to compute the measure is reliable does not seem logically feasible.

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   Panel Member #1 Signal-to-noise analysis and split sample reliability testing.

   Panel Member #2 The developers used 2 different measures of reliability: 1) Reliability score (signal to noise) to evaluate the extent to which variation in the masure is due to true, underlying differences in provider performance (signal) rather than random variation (noise). 2) split-sample reliability testing to examine agreement between 2 scores for a clinician based on randomly-spllit, independent subsets of clinician group practice/clinician episodes. Good agreement indicates the performance score is more the result of clinician characteristics (efficient care) than statistical noise due to random variation. They used 2 years of data (2017-2018) to achieve #'s of episodes per clinician comparable to the numbers used for

actual measurement (35 or more episodes per year) with episodes across years evenly distributed. They used the Shrout-Fleixx interclass correlation coefficient (ICC) between the split-half scores to measure reliability.

Panel Member #3 Reliability Score and Split-sample ICC methodology is acceptable.

Panel Member #4 Signal to Noise ratio, overall and by number of patients. Split -sample correlation. Both appropriate.

Panel Member #7 Reliability scores were a mean of 0.78 and standard deviation of 0.13 for 19,213 TIN's and a mean of 0,70 with a standard deviation of 0.11 for 126,628 TIN-NPI's. Increasing size of practice groups was correlated with an increased reliability score. Split sample intraclass correlation coefficients were 0.66 for TIN and 0.60 for TIN-NPI.

Panel Member #8 Split half; ICC

Panel Member #9 The developer used two approaches to calculate measure score reliability. One is the signal-to-noise reliability with reference to Adams' NEJM paper, another is the split-sample reliability based on Shrout-Fleiss' intraclass correlation coefficient. However, Adams obtained between variance from a two-level hierarchical linear model while this measure is not based on a linear hierarchical lear model, it is not completely clear how different variance components were obtainted to calculate the reliability scores.

7. **Assess the results of reliability testing**

Submission document: Testing attachment, section 2a2.3

Panel Member #1

At a testing volume threshold of at least 35 episodes (the case minimum for the measure in the MIPS 2020 performance period) the mean (SD) reliability for TINs is 0.78 (0.13) and for TIN-NPIs is 0.70 (0.11).

The ICC between the split-sample measures scores for the overall sample were 0.66 and 0.60 for TIN and TIN-NPI, respectively, and may be considered moderate.

Panel Member #2 Overall, reliability testing using 19,213 clinician group practices from 2018 indicated good reliability, regardless of clinician group size. The average reliability score for all clinician group practices was 0.78 with range of 0.67 (25th percentile) to 0.90 (75th percentile). For the 126,628 individual practioners, the mean reliability was slightly lower at 0.70 with range of 0.60 (25th percentile) to 079 (75th percentile) When examined by clinician group size, the average reliability score ranged from 0.70 (1 clinician) to 0.90 (21+ clinicians). The ICC for the overall sample was 0.76 with 95% confidence interval of 0.75-0.77.

The ICC for 17,427 clinican groups as measured by Pearson correlation coefficient was 0.66 and for 95,647 individual practitioners was 0.60. This shows somewhat lower reliability than the signal to noise measure.

Panel Member #3: The Reliability score results were OK; not terrific especially at the individual clinician level measure.

Table 1. Distribution of Reliability Scores for TINs and TIN-NPIs with an Overall Testing Volume Threshold of 35 Episodes

Reporting Level	Number of TINs or TIN NPIs	Mean (Std. Dev.)	25th Pct.	50th Pct.	75th Pct.
TIN	19,213	0.78 (0.13)	0.67	0.79	0.90
TIN-NPI	126,628	0.70 (0.11)	0.60	0.69	0.79

Table 2. Distribution of Reliability Scores for TINs by Practice Size, with an Overall Testing Volume Threshold of 35 Episodes

# of Clinicians	Number of TINs or TIN NPIs	Mean (Std. Dev.)	25th Pct.	50th Pct.	75th Pct.
Overall	**19,213**	**0.78 (0.13)**	**0.67**	**0.79**	**0.90**

1 Clinician	5,771	0.70 (0.10)	0.62	0.69	0.77
2-4 Clinicians	4,022	0.74 (0.10)	0.66	0.74	0.83
5-20 Clinicians	4,739	0.81 (0.11)	0.73	0.83	0.90
21+ Clinicians	4,681	0.90 (0.11)	0.85	0.94	0.98

* Pct. = percentile.

The Split-sample ICC values were even less impressive.

Table 3. Split-sample Intraclass Correlation Coefficients

Reporting Level	# of TINs or TIN NPIs	Mean Score: Sample 1	Mean Score: Sample 2	Pearson Correlation Coefficient	ICC(2,1)
TIN	17,427	1.0132	1.0132	0.66	0.66
TIN-NPI	95,647	1.0412	1.0413	0.60	0.60

Panel Member #4 Reliability is too low for TIN's with 1 or 2-4 Clinicians or TIN-NPI. Mean SN 0.7 and 0.74, with scores at 50th percentile for 1 clinician TINS 0.69, and for 2-4 Clinician practices 0.66. Past experience suggests SN ratios below 0.8 produce unstable estimates.

Split sample ICCs are 0.66 for TIN and 0.60 for TIN-NPI, too low for individual assessment.

Panel Member #7 Reliability scores were a mean of 0.78 and standard deviation of 0.13 for 19,213 TIN's and a mean of 0,70 with a standard deviation of 0.11 for 126,628 TIN-NPI's. Increasing size of practice groups was correlated with an increased reliability score. Split sample intraclass correlation coefficients were 0.66 for TIN and 0.60 for TIN-NPI.

Panel Member #8 Reliability for single clinician and groups of 2-4 clinicians range is low. Developer should consider increasing the volume threshold above 35.

Panel Member #9 Results are good based on two different approaches.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable** (score-level testing was not performed)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☐ **No**

   ☒☐ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and all testing results):

    ☐☒ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☒☐ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

    ☒☐ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

Panel Member #1 Please see 7 above for explanation.

Panel Member #2 **R**eliability testing using two different approaches shows moderate to high reliability, but reliability was considerably lower for smaller agencies (0.63 reliability score and 0.57 ICC for agencies with 20-180 epidoses). Smaller group practices could have much wider variation in performance scores based on sample indicating somewhat lower reliability of the measure.

Panel Member #3 There are several issues regarding how the measures are operationally defined (see comments on exclusions and to whom the clinician measure applies) as well as the moderate at best reliability values. Perhaps if the operational definitions problems are clarified, the rating could move to Moderate.

Panel Member #4 SN for small practices and TIN-NPI and ICC's below level of reliability needed to make assessment at individual level of performance.

Panel Member #5 If I understand the results correctly the reliability testing was performed at the clinician group level; no reliability testing results were reported for at the clinician individual level

Panel Member #7 With the inclusion, exclusion and attribution methodology as stated, the score had moderate reliability.

Panel Member #8 Quartiles of reliability displayed in Tables 1 and 2 range as low as 0.6. This could likely be improved by increasing the volume threshold.

Panel Member #9 Although further clarification on the signal-to-noise approach will be helpful, the results based on split-sample reliability testing are good. Since Shrout's ICC estimate tends to be low, this is reassuring.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

Submission document: Testing attachment, section 2b2.

Panel Member #1 None

Panel Member #2 NONE

Panel Member #3

- Exclusion based on 90-day lookback is confusing. If the episode for determining cost is 3 days before hospitalization to 30 days after, why does Medicare payer status 90 days before the hospitalization matter—and should be excluded?
- Rationale for some exclusions need to be explained/elaborated.

Panel Member #4 Process for evaluating unrelated services for exclusion described but process and system for identifying candidate services to be evaluated is not.

Panel Member #7 Exclusion criteria are provided and when analyzed for impact on cost, did affect the observed cost. It was noted that of 10,658,462 episodes 59.1% remained and had a lower observed cost.

Panel Member #8 40% of episodes are excluded (Table 6) – there is no analysis regarding the level of exclusion by clinician and/or group. This could cause a bias in the measure.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

Submission document: Testing attachment, section 2b4.

Panel Member #1 None

Panel Member #2 NONE

Panel Member #3 The distribution of MSPB values for both the Clinician and Clinician Group scores were normally distributed. The characterization of these scores into three categories as shown in the following table is consistent with these distributions.

Table 7. Proportion of Measure Scores Statistically Significantly Different From the National Average

Provider Level	# of Providers	Statistically significantly lower than national mean		Not statistically significantly different from national mean		Statistically significantly higher than national mean	
		#	%	#	%	#	%
TIN	19,213	3,835	20.0%	12,996	67.6%	2,382	12.4%
TIN-NPI	126,628	19,326	15.3%	97,138	76.7%	10,164	8.0%

Panel Member #4 If I believed the underlying logic of including all services in the post 30 day period, and that the risk adjustment model accurately adjusted for appropriateness of SNF services, then I would say yes, the differences are sufficiently large to be meaningful, but I have questions about both, and therefore don't have confidence in the differences across entitities or O/E that are estimated.

Panel Member #7 Clinician scores do have significant variability with 20% significantly lower than the national mean and 12.4% higher by TIN. For TIN-NPI the numbers are 15.3% and 8.0%, respectively.

Panel Member #9 No concern.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
Submission document: Testing attachment, section 2b5.
Panel Member #1 N/A
Panel Member #2 NONE
Panel Member #3 No comment provided
Panel Member #4 None.
Panel Member #9 No concern.

15. **Please describe any concerns you have regarding missing data.**

Submission document: Testing attachment, section 2b6.

Panel Member #1 None. The missing data being described in section 2b6 of the testing document pertain to the exclusion restrcitions, which have been well-justified

Panel Member #2 NONE

Panel Member #3 There appears to be a large amount of missing data across a large number of both clinicians and clinician groups. This, coupled with the requirement to have at least 35 episodes may limit the reportability of the measure.

Table 8. Missing Data Categories for the MSPB Clinician Measure

Exclusion	# Episodes	# TINs	# TIN NPIs
Missing birth date	*	*	*
Death before trigger	*	*	*
Primary payer other than Medicare	1,132,724	40,140	308,645
Not continuously enrolled in Parts A and B	1,543,418	40,174	328,166

*denotes that there were fewer than 11 episodes

Panel Member #4 None.

Panel Member #7 None as this is Medicare population only.

Panel Member #9 No concern.

16. **Risk Adjustment**

16a. **Risk-adjustment method ☐ None ☒ Statistical model (109 variables) (109 RFs) (Panel Member #3) ☐☒ Stratification by (26 risk categories) (Panel Member #3)**

16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☐☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☒☐ Yes ☐☒ No ☐ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☒ Yes ☐ No

16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☐ No

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒☐ Yes ☐ No X—Not applicable **(Panel Member #3)**

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐☒ Yes ☒☐ No

16d.5.Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☐ No Yes applies to clinical risk factors, NO SES factors were included. (**Panel Member #2**)

16e**. Assess the risk-adjustment approach**

Panel Member #1 The risk adjustment model for the MSPB Clinician measure broadly follows the CMS-HCC risk adjustment methodology used in the Medicare Advantage (MA) program, using an ordinary least squares linear regression model.

Panel Member #2 In spite of making a strong conceptual argument for including SES, statistical results show significant results (t-test and F-test) varied based on p-values indicating that SES factors are predictive for determining resource use for the different stratified DRG groups. They found high correlation of performance scores with and without the risk factors. Due to the inconsistent direction and limited impact of SES effects they concluded the MSPB clinician measure risk adjustment model sufficiently accounts for the effects of SES on measure scores. The developers concluded that adding SES factors, individually or together, did not substantially improve overall model fit. I strongly believe the factors should have been kept in the model. In cases where an HHA serves a large proportion of patients with these SES factors, they will be penalized for having higher costs than expected as these factors are not accounted for, which is not intent of the measure and could restrict access to high quality clinician group practice/clinician for these patients.

The models as specified have good discrimination properties based on clinical risk adjustments applied. The range of R-squared values for the MSPB clinician cost measure risk adjusted models ranges between 0.09 – 0.64 across the MDCs. The adjusted R-squared range is 0.09 – 0.63.

Panel Member #3 Information (results) for the 24 stratifications was never presented. Not clear if there truly was any stratification analyzed.

Panel Member #4 The risk adjustment model is extensive, and has many good characteristics: separate model for each MDC, inclusion of substantial number of disease/illness related indicators, inclusion of indicators for disability and prior institutionalization. However, I am concerned that it underestimates APPROPRIATE use of SNF. This concern is based on data reported in Table 5 Pearson Correlation Statistics between Costs for Clinical Themes with Risk-Adjusted and Expected Costs on page 13 of the testing document. This table shows a 0.06 correlation of SNF care with predicted cost, and 0.34 correlation with risk adjusted costs. By contrast, the correlation of predicted cost for IRF/LTCH is 0.55 and with risk adjusted cost 0.15. Some patients post-hospitalization will require SNF care, even as some SNF care is unnecessary

or excessive. The risk adjuster should therefore adjust out needed or appropriate SNF care. The 0.06 correlation of SNF and predicted cost, while not zero, seems low to me. If there are MDCs with higher expected SNF care and the correlation for those MDCs of predicted costs to SNF costs was higher, this concern might be mitigated.

Panel Member #7 for the MSPB clinician cost measure risk adjusted models ranges MDCs adjusted range is 0.09 – 0.63

Panel Member #9 Risk-adjustment approach is acceptable.

For cost/resource use measures ONLY:

17. **Are the specifications in alignment with the stated measure intent?**

☒ Yes ☒ Somewhat ☐ No (If "Somewhat" or "No", please explain)

Panel Member #4 The intent of this measure is attribute costs to clinicians providing inpatient care the costs of the inpatient care and post-acute care that flow from the hospitalization. Higher O/E costs are taken as an indicator of either excessive services relative to patient need or treatment of complications or other adverse consequences that flow from the care provided by the clinician in the hospital. This requires that the risk adjuster adequately control for higher/lower resource use due to patient condtion and comorbidities, and to appropriate prescription of post-acute services such as SNF, HH, and LTCH. It is not clear that the HCC's and other elements of the risk adjuster robustly differentiate appropriate from inappropriate use of expensive post-acute services.

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):** None

Panel Member #3 See list of concerns about the measure specification, as well as the modest (at best) results from the analyses.

Panel Member #4 Attribution is always an issue in these measures. Some of the downstream decisions on care in the post-acute period may not be attributable to the hospital episode and may not be adequately controlled with the risk adjustment model.

Panel Member #8 Concern regarding the attribution of spend to multiple MSPB calculations. It is unclear which attributed provider (clinician, LTC, hospital, IRF, etc) is responsible for the spend.

Panel Member #9 See my comments on outlier exclusion earlier.

VALIDITY: TESTING

19. **Validity testing level: ☒ Measure score ☐ Data element ☐ Both**

20. **Method of establishing validity of the measure score:**

☒ Face validity

☒ Empirical validity testing of the measure score

☐ N/A (score-level testing not conducted)

21. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

Panel Member #1 Both face validity and empirical validity testing was conducted. Face validity comprised of administering a structured process for gathering detailed input from recognized clinician experts on inpatient care. Multiple expert panels inform the face validity of the measure at different time points: a technical expert panel (TEP), the MSPB service refinement group, and stakeholder feedback from national field testing.

For establishing empirical validity, the measure developer adopted two approaches to empirically examine the extent to which the measure captures what it intends to capture. The first approach sought to confirm the expectation that the MSPB Clinician measure captures variation in service utilization by examining

differences in risk-adjusted cost for known indicators of resource or service utilization (e.g., inpatient readmissions or post-acute care) through the ratio of observed-to-expected cost ("O/E cost ratio").

The second approach empirically tested whether the MSPB measure is capturing variation in provider cost in the manner intended by evaluating how different types of cost arising from homogenous clinical themes (e.g., acute inpatient services, inpatient readmission, post-acute care etc.) impact risk-adjusted measure scores.

Panel Member #2 The developers tested face validity using a structured process to gather input from clinician experts on inpatient care including a technical expert panel an MSPB Service Refinement workgroup, and stakeholder feedback from national field testing.

TEP members completed a face validity survey in November 2019 that assessed (i) the revised measure as compared to the previous version, and (ii) the measure as currently specified after refinements were made. The survey used a Likert scale with values of 1 = Strongly Disagree, 2 = Moderately Disagree, 3 = Somewhat Disagree, 4 = Somewhat Agree, 5 = Moderately Agree, and 6 = Strongly Agree. Fifteen of the 19 TEP members completed the survey

The developers used the following methods to empirically test reliability:

1. Evaluated differences in risk-adjusted cost for known indicators of resource or service utilization, specifically readmission and post-acute care (PAC) utilization. They compared the ratio of observed over expected spending for MSPB-PAC clinician group practice/clinician episodes with and without readmission and with and without PACutilization occurring in the episode period. This analysis tested whether variation in service utilization is captured by the MSPB-PAC cost measure.

2. Empirically tested whether the measure is capturing variation in provider cost by evaluating how different types of cost impact risk-adjusted measure scores. Certain services or costs included in the MSPB Clinician measure were classified into clinically coherent groups of services, called "clinical themes", and are:

- Acute Inpatient Service, including acute inpatient hospital index admission, and services billed by any clinician during index hospitalization
- Inpatient Readmissions, including acute inpatient hospitalization following the index admission and the related services billed by any clinician
- Post-Acute Care (PAC), including home health (HH), skilled nursing facility (SNF), and inpatient rehabilitation or long-term care facility (IRF/LTCH)
- Emergency Services Not Included in a Hospital Admission, including emergency E&M services; procedures; laboratory, pathology, and other tests; and imaging services.
- Outpatient Evaluation and Management Services, Procedure, and Therapy (excluding emergency department), including physical, occupational, or speech and language pathology therapy; E&M services, major procedures; anesthesia, and ambulatory/minor procedures.

They calculated the Pearson correlation between the cost of each clinical theme during the episode and the overall risk-adjusted cost for an episode. Also included are correlations between the cost of each clinical theme and the expected episode spending as predicted by risk-adjustment to show if the resource use within a clinical theme would be expected due to the patient pre-existing conditions outside the influence of the clinician. The hypothesis was that the readmission service category would have the highest correlation with risk-adjusted episode cost, as readmissions are likely associated with high cost even after accounting for beneficiary characteristics. Similarly, they hypothesized that the PAC: SNF service category would have a high correlation with risk-adjusted episode cost since SNF services tend to be less cost efficient than other PAC services (e.g., home health).

Panel Member #3

- The two methods (Face and Empirical) are appropriate for a new measure.
- The methodologies described for these two approaches are appropriate.

Panel Member #4 Face validity was assessed through expert opinion of clinicians, analyzed by survey.

Empirical validity testing relied upon correlation with high cost services, and examining the distribution of O/E ratios for episodes with and without readmission, with and without post acute care (not differentiated by type). There is an extended discussion of why correlation with quality measures could not be adequately conducted.

Panel Member #7 Face validity from expert panel determined the rules for exclusion criteria and attribution logic. Empirical testing was performed by correlation with downstream acute readmission and post-acute care service utilization. Additionally different types of measure scores were grouped into five "clinical themes" and calculation of the Pearson correlation between the cost of each clinical theme during the episode and overall risk-adjusted cost for an episode performed.

Panel Member #9 The developer conducted both face validity and empirical validity testing.

22. **Assess the results(s) for establishing validity**

Submission document: Testing attachment, section 2b2.3

Panel Member #1 Very high level of face validity was confirmed 14 (93%) out of 15 experts agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost effectiveness.

With regard to the first approach in establishing empirical validity, the average O/E cost ratio for episodes with downstream acute readmissions is higher than for episodes without downstream acute readmissions, implying that measure is capable of capturing potentially higher resrouces use due to readmission. Similarly, episodes with PAC services (i.e. HH, SNF, IRF, or LTCH) also have a substantially higher average O/E cost ratio than episodes without PAC services.

The the performance of the second test under empirical validity, the clinical themes analysis, is somewhat lackluster. The test demonstrates somewhat moderate correlation between cost for readmissions and risk-adjusted cost (correlation: 0.47) and weaker correlations with some types of post-acute care services.

Panel Member #2 The results of face validity indicate the experts had a somewhat to moderate level of agreement on average based on survey responses with the measures ability to provide an accurate reflection of costs and distinguish good from poor performance. Out of 15 respondents to the survey, 14 (93%) agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost effectiveness, indicating strong face validity.

Results of empirical validity testing found thhe mean O/E cost ratio for episodes with downstream acute readmission is 1.58, compared with 0.91 for episodes without downstream acute readmission indicating the measure is able to accurately capture higher resource use related to readmission. The mean O/E cost ratio for episodes with PAC is 1.20, while for episodes without PAC is 0.80 as hypothesized.

The results from the clinical themes analysis found weak correlation between the index admission and risk-adjusted cost (correlation: 0.08), but a somewhat stronger correlation between cost for readmissions (0.47) and risk-adjusted cost. Correlation between Outpatient E&M services, procedures, and therapy (0.26) and risk-adjusted cost is low, and correlation between PAC: SNF (0.34) and risk-adjusted cost is moderate, while the correlation between another PAC setting, home health (-0.18) is negative. The correlation between PAC IRF/LTCH (0.15) and risk-adjusted cost was also low.

The positive relationship between MSPB and other indicators of resource/service utilization confirms that the MSPB measure is sensitive to both the occurrence and the intensity of high cost events. The moderate but significant negative correlation between MSPB and DTC measures confirms that, on average, more efficient HHAs are associated with better discharge to community rates and fewer unplanned hospitalizations.

Panel Member #3

- The Face Validity scores were typically 5 out of 6 describing the measures and their results as appropriate.
- The Empirical Validity results showed that the distribution was skewed. In every case the mean value of the comparisons (Measure score vs. downstream/PAC comparator) was higher than the median value. See following table.
- Table 4. Distribution of Observed to Expected Ratios

Cost Driver Category	Observed to Expected Ratios						
	Mean	Std. Dev.	Percentiles				
			10th	25th	50th	75th	90th
All Final Episodes	**1.00**	**0.52**	**0.55**	**0.66**	**0.84**	**1.18**	**1.67**
Episodes with downstream acute (re)admission	1.58	0.66	0.94	1.13	1.41	1.85	2.42
Episodes without downstream acute (re)admission	0.91	0.42	0.53	0.64	0.79	1.02	1.45
Episodes with Post-Acute Care (IRF, LTCH, HH, SN)	1.20	0.56	0.64	0.81	1.06	1.46	1.92
Episodes without Post-Acute Care (IRF, LTCH, HH, SN)	0.80	0.38	0.51	0.60	0.71	0.87	1.14

- The correlation between predicted value and six different "clinical themes" (e.g., PAC settings) were very low (< 0.10) in all cases except PAC IRF/LTCH.
- Correlation with risk adjusted value and six different "clinical themes" also not high—and was negative (-0.18) with PAC HHA.

Panel Member #4 Face validity: The sponsors state "Out of 15 respondents to the survey, 14 (93%) agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost effectiveness.." But the questions they present on face validity on pages 11 and 12 of the testing document ask about: "extent to which you agree that these refinements help the measure provide an accurate reflection of the costs related to inpatient care" [emphasis added] and "provide a more accurate reflection of the costs for inpatient episodes of care than the previous version of the measure…" The questions ask for comparison to earlier versions, not the absolute statement they quote in their summary. Was the question in the summary also asked?

Empirical testing. As noted above, for this measure to be valid, either the service exclusion or the risk adjuster need to adequately control for higher/lower resource use due to patient condtion and comorbidities, and to appropriate prescription of post-acute services such as SNF, HH, and LTCH. The empirical testing does not fully address this issue, since the higher than expected costs with post-acute services. don't fully control for need, and the lumping of the three types of institutional care make analysis of individual components difficult to observe.

Panel Member #7 The mean O/E ratio for all episodes is 1.00 and 1.58 for episodes with downstream readmission versus 0.91 without downstream readmission. The mean O/E cost ratio for episodes with post-acute care was 1.20 veruss 0.80 for those without. Pearson correlation with the clinical themes was most positive for post-acute care SNF and outpatient E&M services, procedures, and therapy.

Panel Member #9 Face validity survery results are very positive, in particular, very high agreement with question 3: The score obtained …. Will provide an accurate reflection of the costs for inpatient episodes of care, and can be used to distinguish good and poor performance on cost effectiveness.

Empirical validity results are as expected. Table 4 presents stratified results by different cost driver categories. Table 5 presents results by clinical theme.

23. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☒ **No** PARTIALLY. REALLY NEED TO SEE A MODEL OF HOW TO DIFFERENTIATE APPROPRIATE FROM INAPROPRIATE OR EXCESSIVE POST ACUTE SERVICES AND OBSERVE THAT THE EXCLUSIONS OR RISK ADJUSTER DO A GOOD JOB OF IMPLEMENTING THAT DIFFERENTIATION. **(Panel Member #4)**

☐ **Not applicable** (score-level testing was not performed)

24. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?**
NOTE that data element validation from the literature is acceptable.

Submission document*: Testing attachment, section 2b1.*

☐ **Yes**

☐ **No**

☒ **Not applicable** (data element testing was not performed)

25. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

☐ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☒ **Low** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

Panel Member #1 Please see my rationale for "moderate" rating in 21 and 22 above.

Panel Member #2 Face validty results showed strong validity related to differentiating episode costs. Empirical validity testing had more mixed results with several measures showing very low correlations to observed/expected risk-adjusted costs.

Panel Member #3 Results from the two methods of establishing validity were not impressive—and reverse relationship between risk adjusted MSPB value and PAC HHA cost suggest that the prediction model is inadequate.

Panel Member #4 Per answer to q22 and q23, for this measure to be valid, either the service exclusion or the risk adjuster need to adequately control for higher/lower resource use due to patient condtion and comorbidities, and to appropriate prescription of post-acute services such as SNF, HH, and LTCH. The empirical testing does not fully address this issue, since the higher than expected costs with post-acute services don't fully control for need, and the lumping of the three types of institutional care make analysis of individual components difficult to observe.

Furthermore, really need to see a model of how to differentiate appropriate from inapropriate or excessive post acute services and observe that the exclusions or risk adjuster do a good job of implementing that differentiation.

The documentation provided on testing was excellent and informative. Documentation on face validity less so.

Panel Member #7 Very dependent on the inclusion/exclusion criteria and attribution criteria. Not sure what the clinical themes adds other than to support that high costs incurred/attributed to clinicians correlates with subsequent high costs.

Panel Member #8 Validity test via stratification by the inclusion of various types of downstream expenses.
Panel Member #9 Outlier exclusion is a key concern. And empirical testing results are to be expected given the functional relationship between the tested components.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

    Panel Member #4 Central to the adoption of this measure is whether the exclusions and risk adjuster adequately differentiate appropriate and efficient levels of post-acute services from inappropriate or unnecessarily high levels of these services. By adopting the standard CMS risk adjustment approach, it is not clear this question was adequately considered.

    Committee also needs to have expert advice on not only what unrelated services were ruled out of the measure but how potentially unrelated services were identified and assessed.

    Panel Member #9 Outlier exclusion should be discussed further. Based on the measure specifications, episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution are excluded. Two factors make this approach particularly concerning: 1) This is a measure focusing on resource utilization, should episodes with very low or very high residuals be excluded? If the concern is with potentially undue influences of outliers, is exclusion the best available approach? 2) Winsorize predicted values: low predicted value below 0.5th percentile is already winsorized before calculation of residual. At minimum, the developer should report the distribution of outlier exclusion across providers to ensure that they don't concentrate in a limited number of providers.

Committee Pre-evaluation Comments:
Criteria 2a: Reliability

2a1. Reliability – Specifications: Describe any additional concerns you have with the reliability of the specifications that were not raised by the Scientific Methods Panel:Describe any data elements that are not clearly defined:Describe any missing codes or descriptors:Describe any elements of the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) that are not clear:Describe any concerns you have about the likelihood that this measure can be consistently implemented:

Comments:

- **N/A**
- **The terms in Step 6 on page 11 need to be more directly linked to the discussion in Step 5, e.g. the "risk-adjusted episode cost ratio" need to be directly defined. Time periods used seem arbitrary and should be condition-specific. Death is endonenous here as poor care results in more death. Analysis should include all patients, and measure should be reported in conjuncture with provider-specific mortality rates across the episodes attributed to a provider. This will enable decision-makers to assess the tradeoffs. Also, "transfer-outs" is a decision of the care provided at the initial hospital and should be included in the measure. "Transfer-ins" are a distinct discussion and seem to be sensibly excluded here.**
- **No additional concerns other than those raised by the Scientific Methods Panel.**
- **Concern about this measure at the level of the individual physician and a few high cost cases leading to instability in the measure at that level. Article by Mehrotra et al. finds measuring cost at clinician level to not be all that reliable. We see this in the reliability stats provided by measure developer. I was not clear (when looking at the excel sheet provided by measure developer) that shows the variables entering the risk adjustment model and the coefficients. It looks like transplants are included. Is that true? I would think these cases are very different from everything else and should be set aside.**
- **No concerns**
- **No concerns that were not raised by the Scientific Methods Panel.**
- **none**
- **No concerns**
- **No concerns**

2a2. Reliability – Testing: Has the developer demonstrated that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers?Describe any additional concerns you have with the reliability testing results or approach that were not raised by the Scientific Methods Panel.

Comments:

- **Share concerns with SMP that S/N and split sample statistics are too low to make assessment at individual level of performance for small practices. Average reliability may be okay, but process seems to have been to pick sample sizes for inclusion before examing reliability rather than making sample sizes consistent with high enough S/N scores.**
- **Yes**

- **None other than those raised by the SMP.**
- **agree with SMP that reliability isn't strong here. A lot of random variation entering into the calculation, especially at clinician level. regarding reliability testing, this statement is very problematic: reporting case minimum have reliability greater than or equal to 0.4, the standard that CMS generally considers as the threshold for 'moderate' reliability. This is incorrect. Anything lower than 0.6 reliability is considered poor and reflects more noise than signal. I have concerns that 50% of all NPIs are below a reliability threshold of 0.69 (between 0.6 and 0.7 is considered low reliability). At a threshold of 35 episodes, this measure doesn't meet reliability for half the physicians being measured, and is weak in discriminating performance. The pearson coefficient for split sample is on the weaker side (of around 0.66) showing disagreement a good share of the time in test/retest of classifying providers.**
- **No concerns.**
- **No concerns that were not raised by the Scientific Methods Panel.**
- **Moderate; would like more clarification on whether reliability was actually tested for clinician individuals (it seems as though the SMP panel wasn't clear on that and it appeared only to reliability test for groups) There is a question of whether the results on reliability for MSBP for clinicians are yield the right results. There are concerns with the lack of information on reliability results below the 25th percentile. CMS states that it considers 0.4 to be the threshold for moderate reliability and that 100% of practices with at least 35 episodes meet it. The minimum acceptable threshold should be higher than 0.4. (some believe it should be 0.7).**
- **No concerns**
- **No concerns**

Criterion 3. Feasibility

3. Feasibility

The extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer states that data are generated by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, medical condition) and are coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims).
- The developer states that all data elements are in defined fields in electronic data sources.
- The developer indicates that there are no fees or licences associated with this measure.

Questions for the Committee:

- *Are there any concerns regarding feasibility?*

Staff preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? Describe your concerns about how the data collection strategy can be put into operational use:Describe any barriers to implementation such as data source/availability, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary tools (e.g., risk adjuster or grouper instrument):

Comments:

- **Measure is routinely compiled from claims and related data, so it is feasible.**
- **Seems feasible**
- **None.**
- **yes, feasible to implement**
- **No concerns with generating the measure, but it will be challenging for clinicians to track and/or improve their performance more regularly.**
- **This measure is feasible to collect and calculate.**
- **Think the availability of social risk data that would improve risk adjustment not yet routinely captured**
- **No concerns**
- **Data elements are routinely generated during care delivery**

Criterion 4: Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

Use

4a.1. Accountability and Transparency.

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4a.2. Feedback on the measure by those being measured or others.

Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure.

4a1. Current uses of the measure

- **Publicly reported?** ☒ **Yes** ☐ **No**
- **Current use in an accountability program?** ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- The developer indicates that this measure is used within the Quality Payment Program Merit-based Incentive Payment System (https://qpp.cms.gov/mips/overview).
- The developer states that, as specified in the CY 2020 Physician Fee Schedule final rule (84 FR 62959 through 62979), this measure will be implemented as part of MIPS beginning in the 2020 MIPS performance year and 2022 MIPS payment year.

4a2.Feedback on the measure by those being measured or others

- The developer collected feedback during the development and implementation of the measure and provided education and outreach through webinars and email communications.
- The developer notes that the overarching feedback that was received on measure performance and implementation from the measured entities and others included comments that (i) the revised specifications made several improvements to the measure; (ii) while field test reports and other supplementary materials were helpful, the complexity of these documents was a challenge to some stakeholders; and (iii) general questions on and proposed updates to the measure's attribution methodology.

Additional Feedback:

- This measure was reviewed by the Measure Applications Partnership (MAP) for 2018 – 2019 meausures under consideration.
- The MAP conditionally supported this measure pending NQF endorsement. The MAP noted that this measure would be an update to the existing MSPB measure in MIPS but noted that neither the updated nor the original measure has been reviewed by NQF Cost and Efficiency Standing Committee, limiting the ability of the group to determine the validity of the changes to the measure. MAP noted specific considerations for this measure. Specifically, MAP urged CMS to continue testing the primary changes to this measure, which are removing costs that are unlikely related to the clinician and a new attribution model, to ensure that they produce the intended results. In particular, MAP noted the need to ensure the measure demonstrates validity and reliability at the NPI level. MAP also noted the desire to avoid double counting clinician costs in the total cost measures and the episode-based cost measures and for CMS to consider consolidating the MSPB and TPCC measures to avoid overlap. MAP also expressed concern about the challenges of getting access to field test data. MAP encouraged CMS to monitor for unintended consequences to patients such as under treatment, impact on technology innovation, and access to treatment for high-risk, high-resource use patients.
- Lastly, MAP urged CMS to continuously test and refine the risk adjustment model and incorporate social risk factors when appropriate. MAP also recommended that QIOs assist in providing education on this measure to clinicians.

***Questions for the Committee**:*

- *Does the Committee have any concerns with its use?*

Staff preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

Usability

4b.1 Improvement.

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible

rationale describes how the performance results could be used to further the goal of high quality, efficient healthcare for individuals or populations.

4b2. <u>Benefits vs. harms</u>.

Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b3. Data and result detail are maintained such that the resource use measure, including the clinical and construction logic for a defined unit of measurement, can be deconstructed to facilitate transparency and understanding.

4b1. Improvement results

- This measure is being considered for initial endorsement.

4b2. Unintended consequences

- The developer states that there are no unexpected findings during the development and testing fo the measure.

4b2.Potential harms

- The developer states that there are no unexpected findings (including harms and benefits) during the development and testing fo the measure.

4b3. Transparency

- This measure is publicaly reported.
- Measure specifications and methods are transparent and available for users.

***Questions for the Committee**:*

- *What benefits, potential harms, or unintended consequences should be considered?*
- *Do the measure specifications and accompanying documentation enable adequate transparency to facilitate understanding of how the measure results are generated?*

Staff preliminary rating for Usability: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency: How is the measure being publicly reported?Is the measure being used in any other accountability applications?Are the performance results disclosed and available outside of the organizations or practices whose performance is measured?Is a credible plan for implementation provided?

Comments:

- **Being used in MIPS payment program.**
- **Not sure**
- **Yes. See the information provided by the NQF staff to this (4a1).**
- **It is an overall summary measure, however, absent CMS providing drill down information, it is hard for clinicians and practices to determine where to focus to make changes. CMS is working to implement the measure as part of MIPS CY 2020. This will be part of the MIPS payment method (incentives). CMS is not yet publicly reporting the measure but may in the future.**
- **No comment.**
- **This measure is not publicly reported but is shared with MIPS participants.**
- **yes; Theoretically – as data files are made available. Unclear whether clinicians/groups learn which patients have shared attribution and which clinicians/groups they are sharing responsibility with (in order to develop coordination plans)**
- **No concerns**
- **Planned use**

4a2. Use – Feedback: Describe any concerns with the feedback received or how it was adjudicated by the measure developer: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data?Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation?Has this feedback has been considered when changes are incorporated into the measure?

Comments:

- **yes**
- **See thoughts above.**
- **Satisfied with the summary on the above questions compiled by the NQF staff (see section 4a2).**
- **yes, this measure has been vetted through TEP, the working group and through testing with providers**
- **n/a**
- **Unclear. Report cites MAP review but not clinician feedback on the measure performance/implementation**
- **No concerns**

- **Planned use**

4b1. Usability – Improvement: Has the measure developer demonstrated that the use of this measure is helping to drive improvements in cost or efficiency?Has the developer adequately described how the performance results be used to further the goal of high-quality, efficient healthcare?If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations?

Comments:

- **I am not clear on what is being reported as part of the measure back to providers, including whether individual cost components and relative performance by cost component is provided.**
- **Seems like episode windows should be condition-specific.**
- **The developer has provided evidence that the implementation of this measure will reduce spending variability across Medicare physician or physician groups with regard to MSPB.**
- **As a summary measure it provides a broad gauge to attributed physician re: extent to which their episode spending is greater than others. However, it lacks the drill down specificity to help clinicians/TINs quickly identify where the key difference areas are that are driving overall differences in spending to help with QI efforts. As such the burden is on the provider to try to sort this out.**
- **n/a**
- **No concerns**
- **Planned use**

4b2. Usability – Benefits vs. harms: Describe any unintended consequences and note how you think the benefits of the measure outweigh them:

Comments:

- **Low reliability for small practices suggest they could be subject to windfall gains and losses for care variations not under their control.**
- **Dropping "death episodes" is problematic and could lead to misleading performance inferences.**
- **None.**
- **Not adjusting for social risk factors runs the risk of providers avoiding these patients. this is real given that this measure score constitutes 15% of the overall measure weight in the MIPS score and $$ are attached to it. It has the potential to further under-resource those providers who need more resources to care for more complex (social risk factor) patients.**
- **n/a**

- **access to appropriate care is a significant concern and unintended consequence/potential harm.**

Developer simply states no unexpected findings for harm or unintended consequences. Costs should be assessed within the context of the quality of care provided. Yet the developer does not demonstrate that this measure correlates to any of the quality measures within the QPP. The developer should consider assessing the MSPB clinician measure with a measure, such as the claims-based All-Cause Hospital Readmissions, which was also reported in 2017, and was attributed to practices that same year.

- **No concerns**
- **No concerns**

Criterion 5: Related and Competing Measures

- There are no competing measures (i.e. same measure focus and target population)
- The developer states that this measure is related to measure(s):
    - 2158 : Medicare Spending Per Beneficiary (MSPB) – Hospital

Harmonization

- The developer states that "the MSPB Hospital and MSPB Clinician measures are closely aligned and that both measures assess costs from the same time window (three days prior to the index admission to 30 days after discharge) and focus on the same target population of beneficiaries admitted to the inpatient setting."

Committee Pre-evaluation Comments: Criterion 5:
Related and Competing Measures

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

Comments:

- **As noted in documentation, parallel measure for hospitals. Developer views as complementary.**
- **No competing measure identified. Related NQF-endorsed measure identified by the developer is: 2158: Medicare Spending Per Beneficiary (MSPB) – Hospital. The developer stated that the MSPB Hospital and MSPB Clinician measures are closely aligned and that both measures assess costs from the same time window.**
- **other related spending measures. none competing.**
- **I'm not sure if harmonizing is required, but at a minimum, better articulating and distinguishing between the MSPB Clinician measure (#3574), HealthPartner's Total Cost of Care measure, and CMS' Total Per Capita Cost (#3575) measure would be helpful.**
- **none.**
- **Interaction with the TPCC measure not discussed by the developer. Developer raises the Hospital MSPB measure is aligned as they both measure from the same time window and focus on the same target patient population. Unclear whether there's any effort to prospectively share attribution information between clinicians/groups and hospitals to assist care coordination and discussions of appropriate use. The MSPB measure captures the same costs as the episode-based measures, effectively "double counting" the costs when both are used in the MIPS program.**
- **No concerns**
- **No concerns**

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: July 1, 2020

- Comment by American Academy of Neurology (AAN)
  Member Vote N/A
  8408 :
  The American Academy of Neurology (AAN) appreciates the opportunity to comment on this measure and hopes the Cost and Efficiency Technical Advisory Panel takes these comments into consideration when deliberating on the measure.

  The AAN echoes the American Medical Association's concerns related to the measure score reliability, empirical validity and risk adjustment methodology for this measure. While The Centers for Medicare and Medicaid Services (CMS) developed the Medicare Spending Per Beneficiary (MSPB) Clinician measure for use in the Merit-based Incentive Payment System (MIPS) to meaningfully and reliably distinguish individual and groups by measuring costs associated with the care provided to patients, the measure testing results are unclear and fail to demonstrate reliable and valid results that support moving forward with measure endorsement and implementation at this time. Our top concerns based on the measure testing results include:

- An inadequate moderate reliability threshold
- A lack of correlation to quality measures used in MIPS; cost of care assessments should be rooted within the context of quality measure assessment, which this measure fails to do
- An inadequate and unclear risk adjustment model, including lack of appropriate testing and adjustment for social risk factors

With these concerns in mind, the AAN does not support the measure based on the testing results provided and these gaps should be addressed before endorsement and implementation.

- Comment by American Psychiatric Association (APA)
  Member Vote N/A
  8407 :
  The American Psychiatric Association appreciates the opportunity to submit comments for the Cost and Efficiency Standing Committee's review. APA continues to have serious concerns about the Medicare Spending Per Beneficiary (MSPB) measure, and concurs with the American Medical Association (AMA)'s more detailed analysis of the measure.

  It is not clear that clinicians can control the costs that are attributed to them as part of this measure, particularly those costs that are incurred after hospital discharge. In addition, the developer notes that they are unable to adequately test the relationship between performance on the cost measure and performance on conceptually-related quality measures, including patient outcomes. This is a relationship that should be explored more thoroughly before implementation of the measures, to guard against unintended consequences or mis-alignment of incentives for healthcare providers.

- Comment by Infectious Diseases Society of America (IDSA)
  Member Vote N/A
  8404 :
  IDSA appreciates the opportunity to provide comments to the NQF Cost and Efficiency Standing Committee. IDSA agrees with the findings of the American Medical Association's (AMA) more detailed analysis of the MIPS Medicare Spending per Beneficiary (MSPB) and Total per Capita Cost (TPCC) measures. We continue to have concerns about the ability of these measures to accurately and reliably distinguish performance among clinicians, the ongoing failure of these cost measures to link to relevant quality measures under MIPS, and the ongoing failure of these measures to produce meaningful and comprehendible information that clinicians can use to enhance patient care and value. We are also concerned that ID physicians may be held accountable simultaneously for both cost measures under MIPS. While recent revisions to these measures were intended to avoid this situation, many members of our specialty work in both inpatient and outpatient settings. As a result, they may be captured under the MSPB measure under the medical E/M attribution rule, but also under the TPCC measure since the ID specialty is not specifically excluded from this measure.

- Comment by American Society of Clinical Oncology (ASCO)
  Member Vote N/A
  8402 :
  The American Society of Clinical Oncology (ASCO) appreciates the opportunity to submit comments to the National Quality Forum (NQF) Cost and Efficiency Technical Advisory Panel. Following are our general comments on the Medicare Spending per Beneficiary (MSPB) and Total per Capita Cost (TPCC) measures.

  ASCO is the national organization representing nearly 45,000 physicians and other health care professionals specializing in cancer treatment, diagnosis, and prevention. ASCO members are also dedicated to conducting research that leads to improved patient outcomes, and we are committed

to ensuring that evidence-based practices for the prevention, diagnosis, and treatment of cancer are available to all Americans, including Medicare beneficiaries.

Given the growing number of episode-based cost measures, and continued work on their development, ASCO would encourage the NQF and CMS to consider whether the TPCC and MSPB measures still serve a purpose, as many of the beneficiaries captured in the episode-based measures will also be included in either or both the MSPB and TPCC measures. With the measures as proposed, a beneficiary could potentially be attributed to multiple providers within and across multiple measures. First, this could magnify the impact on cost measures of any individual beneficiary and second, could complicate any true differences in cost and value. CMS developed these measures specifically for use in the Merit-based Incentive Payment System (MIPS) and we believe that the measure and attribution should demonstrate that its use in MIPS will not just yield reliable and valid results, but most importantly, enable end users to make meaningful distinctions in the costs associated with the care provided to these patients.

- Comment by American College of Physicians (ACP)
  Member Vote N/A
  8397 :
  The ACP appreciates the opportunity to comment in advance of the NQF Cost and Efficiency Standing Committee's review of several measures submitted for endorsement consideration during the Spring 2020 cycle.

  The Medicare Spending per Beneficiary (MSPB) measure represents an important move towards cost assessment in pay-for-performance programs. However, the methods that policymakers and measure developers apply to assessing episode-based costs is critical to the success of this initiative. In this regard, several inherent limitations to the measure exist. The Centers for Medicare and Medicaid Services (CMS) should consider addressing the concerns listed below in the interest of enhancing the validity of the measure.

  The Performance Measure Committee (PMC) of the ACP prefers that all cost measures be attributed to the level of the group/practice or higher for the following reasons:

  If health plan administrators and government payers intend to create individual cost profiles to generate incentives to decrease health care costs, it is important that these profiles provide insights into which care management interventions are most effective in reducing costs year-over-year, even if what is measured does not encompass the totality of the cost to Medicare for the items and services provided to a patient during an episode of care. Measuring what is actionable could build trust with clinicians, feed a cycle of participation, and discourage dysfunctional behaviors such as avoiding attribution. Stratifying and comparing results based on costs related to 1) services that are under the direct control of the individual clinician, 2) indirect costs, and 3) services under the control of the facility could help to mitigate this concern by identifying behaviors that correspond with opportunities for improvement.

  While improvements have been made to the attribution model, revisions do not address the possibility of multiple clinicians being held accountable for the total costs associated with a single episode. CMS attributes each MSPB episode to the Taxpayer Identification Number-National Provider Identifier (TIN-NPI) responsible for 30% of Part B Physician/Supplier services during the index admission.

  According to this model, multiple clinicians could be accountable for the total costs associated with a single episode of care. While we generally support the attribution model at the facility, system, and health plan levels, we caution that attributing patient costs to individual clinicians can be

technically challenging. Healthcare costs are influenced not only by the actions of one clinician but often by the actions of multiple clinicians as well as a patient's social, economic, and environmental factors. It is difficult to determine the relative influence that an individual clinician has on a patient's expenses. Understanding who is responsible is essential to driving improvements in care as well as for securing long-term buy-in from clinicians and facilitating the ability of value-based purchasing programs to influence clinician behavior. The current model does not speak to the care coordination system that most clinicians would likely endorse. For example, Accountable Care Organizations that build on the value-based purchasing framework to enhance care coordination and promote responsibility for clinical and efficiency outcomes.

Additional areas of concern are as follows:

We are unable to assess the benefit of assessing costs (e.g., if it helps to improve outcomes at lower costs) without assessing the evidence to support this claim. We recommend that NQF require that measure developers document the evidence base for cost/resource use measures and that it at least aligns with what is required for outcome measures (i.e., Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service).

The implications of the risk-adjustment model as currently specified are unclear. The model estimates expected episode costs in recognition of the different levels of care beneficiaries may require due to comorbidities, disability, age and other risk factors. This model is not sufficient to control for all significant social determinants of health (SDOH) that may influence the clinical health status of patients as well as the outcome of acute admissions. The Centers for Medicare and Medicaid Services (CMS) should consider revising the risk-adjustment model to include SDOH that are most likely to influence the clinical health status of the denominator population under consideration. Aligning the model for risk-adjustment with more robust methods for statistical analyses that consider all factors that are independently and significantly associated with outcomes and that vary across measurement participant (e.g., the Society for Thoracic Surgeons Adult Cardiac Surgery Risk Model) could enhance individual clinician acceptance of outcomes measures and helps to mitigate risk aversion.

The 30% threshold is too low to attribute episode-based care to an individual clinician. CMS should consider increasing the attribution threshold to an evidence-based percentage that represents the majority of services during hospitalization.

The 30-day episode window is arbitrary. Recent literature suggests that shorter intervals of seven or fewer days might improve the accuracy and equity of episode-based costs to Medicare as a measure of facility quality for public accountability.

While we note that the current use of this measure requires that clinicians and clinician groups meet a 35-episode case minimum which is referenced in a few sections of the submission form, we would recommend that this minimum requirement be included in the technical measure specifications - either in the denominator requirements or exclusions. This is particularly important given that the measure's reported reliability results rely on a minimum volume threshold of 35 episodes.

Maximizing transparency could build trust with clinicians and feed a cycle of participation. CMS should consider establishing a premortem approach for evaluating the impact of performance measures to combat the unintended consequences of implementation and correctly identify reasons for future outcomes.

While this measure aims to reduce low-value care, implementation may result in consequences directly contrary to the spirit of the measure. The measure specifies "episodes of care for a beneficiary if the beneficiary dies during the episode" as exclusion criteria. Therefore, the measure rewards clinicians for expending minimal resources on patients in stable conditions, while disregarding mortality rates, and penalizes clinicians for disbursing sufficient resources to maintain the stability of medically complex patients during an episode of care.

- Comment by American Association of Neurological Surgeons (AANS)
  Member Vote N/A
  8395 :
  The American Association of Neurological Surgeons (AANS) thanks the NQF for the opportunity to share input on the MSPB Clinician measure (#3574), which was developed and recently revised for use under the Merit-Based Incentive Payment System (MIPS). As noted by the American Medical Association's (AMA) more in-depth analysis, the information and testing provided—particularly for measure score reliability, empirical validity and the risk-adjustment approach— do not demonstrate that the use of this measure under MIPS will yield reliable or valid results or enable us to distinguish low versus high performers. As a result, end users will not be able to make meaningful distinctions regarding the costs associated with the care provided to these patients. Furthermore, the AANS has long voiced concerns about this measure's failure to evaluate cost in the context of quality. The revised version of this measure still does not correlate to any one quality measures used in MIPS.

  Given these ongoing concerns, the AANS respectfully requests that the NQF not endorse this measure until these deficiencies have been addressed by CMS.

- Comment by American Medical Association (AMA)
  Member Vote N/A
  8389 :
  The American Medical Association (AMA) appreciates the opportunity to comment on this measure and request that the Standing Committee carefully consider our comments on its scientific acceptability during this evaluation.

  The Centers for Medicare and Medicaid Services (CMS) developed this measure specifically for use in the Merit-based Incentive Payment System (MIPS) and we believe that the information and testing provided should demonstrate that its use in MIPS will yield reliable and valid results and enable end users to make meaningful distinctions in the costs associated with the care provided to these patients. The AMA is concerned that the testing results provided, particularly for measure score reliability, empirical validity and the risk adjustment approach, do not provide the information needed to ensure that MSPB Clinician produces the desired results.

  Regarding the measure score reliability, we are concerned with the lack of information on reliability results below the 25th percentile, particularly in light of the reference within the response of 2a2.3 that CMS generally considers 0.4 to be the threshold for moderate reliability and 100% of practices and clinicians with at least 35 episodes meet it. The AMA believes that the minimum acceptable thresholds should be 0.7 and the measure as specified does not.

  The AMA strongly supports the tenet that cost must be assessed within the context of the quality of care provided; yet, the developer was unable to demonstrate that this measure correlates to any one quality measure within the MIPS program and differs from what they were able to complete for other MSPB measures currently under review (3561, 3562, 3563, and 3564). We are very troubled that the testing did not include an assessment of MSPB Clinician with a measure such as the claims-based All-Cause Hospital Readmissions (#458) since it was also reported in 2017 and to our knowledge CMS attributed performance to practices for which this cost measure could also apply in

that same year. While we acknowledge that the lack of alignment of attribution models creates challenges to complete these analyses, we believe that CMS could solve this issue since the agency serves as the steward for many of the claims-based measure. Regardless, the AMA does not believe that cost measures against which no quality measure can be assessed should achieve endorsement.

The AMA does not believe that the current risk adjustment model is adequate due to R-squared results ranging from 0.09 to 0.64 across the groupings nor is the measure adequately tested and adjusted for social risk factors. It is unclear to us why the developer would test social risk factors after adjusting for clinical risk factors rather than assessing the impact of both clinical and social risk factors in the model at the same time. These variations in how risk adjustment factors are examined could also impact how each variable (clinical or social) perform in the model and remain unanswered questions. In addition, while the developer believes that the small differences in measure results "can be interpreted as meaningful" (response in 2b4.2 in the testing form), it is not clear why this same reasoning was not applied for those clinicians and practices for whom inclusion of social risk factors in the models changed the ratios nor is it clear how these same factors would affect a change in performance across the 10 deciles used in the MIPS benchmarking methodology.

In addition, the AMA questions whether the information provided in Section 2b4. Identification of Statistically Significant and Meaningful Differences in Performanceis truly useful for accountability and informing patients of the cost of care provided byphysicians and practices. Specifically that the testing does not directly address whether the costs attributed to physicians and practices enable us to distinguish low versus high performers. Since this measure was specifically developed for use in MIPS, analyses of the performance scores using the finalized benchmarking methodology across 10 deciles would provide valuable information on whether the differences in costs between physicians and practices could be considered useful and meaningful. We do not believe that stratifying scores by characteristics such as region, risk score, or the number of episodes attributed satisfactorily answers this question.

The AMA requests that these gaps in testing be addressed prior to endorsement of this measure. We appreciate the Committee's consideration of our comments.

Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

IM.1. Opportunity for Improvement

IM.1.1. Briefly explain the rationale for this measure (e.g., the benefits or improvements in performance envisioned by use of this measure)

Increases in Medicare spending have been an important driver of rising total health care expenditures in the United States. [1] Given that the inpatient hospital setting is a significant contributor to overall Medicare spending, gauging the efficacy of this spending requires measuring the cost performance of clinicians providing care at hospitals. [2] The MSPB Clinician measure provides valuable context for such progress by measuring costs of care from a holistic perspective at the beneficiary level and offering a tool to control rising health care costs. [3]

[1] "National Health Expenditure Projections, 2017-2026." US Centers for Medicare & Medicaid Services, 2018.

[2] "Report to the Congress: Medicare Payment Policy." MedPAC, 2018. http://www.medpac.gov/docs/default-source/reports/mar18_medpac_entirereport_sec.pdf.

[3] Data Book: Health Care Spending and the Medicare Program." MedPAC, 2017. http://www.medpac.gov/-documents-/data-book.

IM.1.2. Provide performance scores on the measure as specified (current and over time) **at the specified level of analysis.** (This is required for endorsement maintenance. Include mean, stddev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include). **This information also will be used to address the subcriterion on improvement (U.3.1.) under Usability and Use.**

Performance scores are provided for 19,213 clinician group practices (identified by Tax Identification Number [TIN]) and 126,628 practitioners (identified by a combination of TIN and National Provider Identifier [NPI]). These counts represent attributed clinicians and clinician groups billing Part B Physician/Supplier claims under a Merit-based Incentive Payment System (MIPS)-eligible clinician specialty, and do not reflect other MIPS eligibility criteria (e.g., Advanced APM participation). Clinicians and clinician groups are included if they are attributed 35 or more MSPB Clinician episodes, as identified in Medicare Parts A and B claims data, during January 1, 2018, to December 31, 2018. Episodes from all 50 States and D.C. with an index admission in the acute inpatient setting were included.

TIN Level Scores:

- Mean score: $19,194
- Standard deviation: $1,785
- Min score: $10,134
- Max score: $36,432
- Score IQR: $2,049
- Score percentiles

o 10th: $17,152

o 20th: $17,845

- 30th: $18,311
- 40th: $18,694
- 50th: $19,055
- 60th: $19,427
- 70th: $19,873
- 80th: $20,453
- 90th: $21,385

- Number of beneficiaries: 4,114,268

TIN-NPI Level Scores:

- Mean score: $19,741
- Standard deviation: $1,885
- Min score: $11,020
- Max score: $39,646
- Score IQR: $2,335
- Score percentiles
  - 10th: $17,504
  - 20th: $18,233
  - 30th: $18,754
  - 40th: $19,198
  - 50th: $19,632
  - 60th: $20,071
  - 70th: $20,566
  - 80th: $21,163
  - 90th: $22,067
- Number of beneficiaries: 3,755,580

IM.1.3. If no or limited performance data on the measure as specified is reported in IM.1.2., then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

N/A.

IM.1.4. Provide disparities data from the measure as specified (current and over time) **by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** (This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) **This information also will be used to address the subcriterion on improvement (U.3.1.) under Usability and Use.**

N/A.

IM.1.5. If no or limited data on disparities from the measure as specified is reported in IM.1.4., then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.

N/A.

IM.2. Measure Intent

IM.2.1. Describe intent of the measure and its components/ Rationale (including any citations) for analyzing variation in resource use in this way.

As background to the MSPB Clinician measure, a version of the measure (referred to as the "MSPB measure") was originally used in the Physician Value-Based Modifier (VM) Program and reported in the annual Quality and Resource Use Reports (QRURs). With the introduction of the Quality Payment Program, the MSPB measure was finalized with minor adaptations from the VM Program's version and added to the Merit-based Incentive Payment System (MIPS), where it was part of the MIPS cost performance category during the 2017-2019 MIPS performance periods. In 2018, the MSPB measure went through re-evaluation to address stakeholder feedback received from prior public comment periods. This stakeholder input informed modifications to the MSPB measure's attribution methodology and included service exclusions. The resulting MSPB Clinician measure will be used in MIPS starting with the 2020 performance period. A summary of the differences between the submitted MSPB Clinician measure for use in the MIPS 2020 performance period and the previously used version of the MSPB measure can be found in Appendix B of the Measure Information Form for the MSPB Clinician measure on the CMS MACRA Feedback webpage. [1]

Rationale for Measuring Cost of Medicare Spending Per Beneficiary (MSPB) Clinician

A recent study indicates that physician beliefs about treatment may be the most important factor explaining the variation in health care expenditures. [2] However, these same clinicians are often unaware of how their care decisions can influence the overall costs of care. One of the goals for using cost measures is to help inform clinicians of the cost of their patient's care, as well as provide detail that is informative and actionable for clinicians. Clinicians may be able to review these costs and determine which are most high yield and efficient.

As health expenditures continue to increase in the United States, the MSPB Clinician measure is an important means of measuring Medicare spending, which is the largest single purchaser of health care in the United States. According to the National Health Expenditure Accounts, total health care spending is estimated to have increased by 4.6 percent in 2017, reaching $3.5 trillion, and spending for Medicare, which is still predominantly paid on a fee-for-service (FFS) basis, grew by 3.6 percent, reaching $672.1 billion. [3] In 2016, Medicare FFS paid $183 billion for approximately 10 million Medicare inpatient admissions and 200 million outpatient services, which reflects a 2.3 percent increase in hospital spending per FFS beneficiary between 2015 and 2016. [4] Inpatient hospital spending, specifically, is an important contributor to overall costs, accounting for 22 percent of total Medicare spending in 2015 and representing the second largest Medicare spending category in 2015 . [5]

Successfully establishing payment models under MIPS can have significant impacts on reducing costs and making care more affordable. [6] Population-based measures serve an essential role in measuring the cost of care and can serve as a transparent tool to control and curb growing health care costs. The MSPB Clinician measure can provide valuable context for such progress in quality, by clarifying the simultaneous movement in the average costs of hospital admissions. By risk-adjusting episode costs, the MSPB Clinician measure provides more actionable information to clinicians and policymakers than a simple trend in overall spending, since the latter does not account for patients' severity of illness or other factors that can affect the costs of an admission.

Rationale for Use of Claims Data to Measure Cost

- There is no additional submission burden, as clinicians must already submit claims for reimbursement.
- Using Medicare Parts A and B claims data allows CMS to evaluate TIN and TIN-NPI cost across all conditions and procedures, resulting in a comprehensive set of data on MSPB Clinician cost performance.
- Additionally, the wide reach of Medicare claims data maximizes the impact of the measure, ensuring that the most TINs and TIN-NPIs benefit from the information provided on MSPB Clinician cost performance.

[1] CMS, "Merit-Based Incentive Payment System (MIPS): Medicare Spending Per Beneficiary (MSPB) Clinician – Measure Information Form," https://www.cms.gov/files/zip/2020-cost-measure-information-forms.zip.

[2] Source: Cutler, D., Skinner, J.S., Sterm. A.D., Wennberg, D."Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending" American Economic Journal: Economic Policy 11(1), 192-221, Feb 2019)

[3] "National Health Expenditure Projections, 2017-2026." US Centers for Medicare & Medicaid Services, 2018.

[4] Report to the Congress: Medicare Payment Policy." MedPAC, 2018. http://www.medpac.gov/docs/default-source/reports/mar18_medpac_entirereport_sec.pdf.

[5] Data Book: Health Care Spending and the Medicare Program." MedPAC, 2017. http://www.medpac.gov/-documents-/data-book.

[6] Ibid.

Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

De.6. Non-Condition Specific *(check all the areas that apply):*

De.7. Care Setting *(Select all the settings for which the measure is specified and tested):*

Inpatient/Hospital

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

https://qpp-cm-prod-content.s3.amazonaws.com/uploads/812/2020+MIPS+Cost+Measure+Info+Forms.zip and https://qpp-cm-prod-content.s3.amazonaws.com/uploads/811/2020+MIPS+Cost+Measure+Code+List.zip

S.2. Type of resource use measure *(Select the most relevant)*

Per episode

S.3. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED):*

Clinician : Group/Practice, Clinician : Individual

S.4. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

S.5. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.5.1.

Assessment Data

Claims

Enrollment Data

Other

S.5.1. Data Source or Collection Instrument *(Identify the specific data source or data collection instrument, e.g. name of database, clinical registry, collection instrument, etc.)*

Medicare Part A and Part B claims data: Part A and B claims data are used to build MSPB Clinician episodes, calculate episode costs, and construct risk adjustors. CMS Office of Information Systems (OIS) maintains a

detailed Medicare Claims Processing Manual available at the following URL: https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Internet-Only-Manuals-IOMs-Items/CMS018912.

Medicare Enrollment Database (EDB): This is used to determine beneficiary-level exclusions and supplemental risk adjustors, specifically Medicare Parts A, B, and C enrollment; primary payer; disability status; end-stage renal disease (ESRD); beneficiary birth dates; and beneficiary death dates.

Minimum Data Set (MDS): The MDS is used to create the Long Term Care Indicator variable in risk adjustment. Data documentation for the MDS is available at the following URL: https://www.resdac.org/cms-data/files/mds-3.0.

We used additional data sources for measure testing purposes:

•American Community Survey (ACS): This is used for evaluating social risk factors. https://www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html.

•Common Medicare Environment (CME) database: This is used for evaluating social risk factors. https://www.ccwdata.org/documents/10280/19002256/medicare-enrollment-impact-of-conversion-from-edb-to-cme.pdf.

S.5.2. Data Source or Collection Instrument Reference *(available at measure-specific Web page URL identified in S.1 OR in the file attached here) (Save file as: S_5_2_DataSourceReference)*

<SamplingMethodologySpecificDataSourceAttachment nodeType="0">2020_01_06_testing_form_appendix_mspb_clinician.xlsx

S.6. Data Dictionary or Code Table *(Please provide a web page URL or attachment if exceeds 2 pages. NQF strongly prefers URLs. Attach documents only if they are not available on a web page.)*

Data Dictionary:

URL: The Research Data Assistance Center (ResDAC) maintains Medicare claims and administrative data dictionaries: https://www.resdac.org/file-availability-vrdc CMS maintains the Medicare Enrollment Database and data dictionary: edbonline@cms.hhs.gov

Please supply the username and password:

Attachment:

Code Table:

URL:

Please supply the username and password:

Attachment: 2020-04-29-mspb-clinician-measure-codes-list-icd-10-codes.xlsx

Construction Logic

S.7.1. Brief Description of Construction Logic

If applicable, summarize the general approach or methodology to the measure construction. This is most relevant to measures that are part of or rely on the execution of a measure system or applies to multiple measures.

The MSPB Clinician measure is a clinician's average risk-adjusted cost across all episodes attributed to the clinician (TIN-NPI) or clinician group (TIN). The measure population is defined by admission to an inpatient hospital. The episode window starts 3 days prior to this index admission and ends 30 days after discharge. Episodes are attributed to clinicians and clinician groups based on the Part B services provided during a medical or surgical Medicare Severity Diagnosis-Related Group (MS-DRGs) inpatient admission. The costs of all services occurring during the episode window - except for a limited set of services that are not clinically related to the management of care for the episode - are summed to obtain each episode's standardized

observed cost. A regression model is applied to the risk adjustment variables to estimate the expected cost of each episode.

The cost measure is calculated as a the ratio of standardized observed cost to expected cost averaged across all of a clinician or clinician group's attributed episodes to obtain the average episode cost ratio. The average episode cost ratio is multiplied by the national average observed episode cost to generate a dollar figure for the cost measure score.

S.7.2. Construction Logic *(Detail logic steps used to cluster, group or assign claims beyond those associated with the measure's clinical logic.)*

STEP 1: Define and Trigger Episodes

Episodes are opened, or triggered, by admissions to inpatient hospitals. The episode window starts 3 days prior to this index admission and ends 30 days after the hospital discharge. There is a 90-day lookback period before the episode start date. This period is used to check beneficiary enrollment information for episode exclusions and beneficiary pre-existing health characteristics used for risk adjustment.

STEP 2: Attribute Episodes to Clinicians

Attribution is the process of determining which clinician groups and clinicians are responsible for an episode. The MSPB Clinician measure utilizes two attribution methods for medical and surgical MS-DRG episodes:

•Medical episodes (i.e., episodes for which the index admission has a medical MS-DRG) are attributed to any clinician/clinician group responsible for managing the medical condition during the inpatient stay. Specifically, the episode is attributed first to the TIN that bills at least 30 percent of E&M codes found on Part B Physician/Supplier claims during the inpatient stay. The episode is then attributed to the TIN-NPI who billed at least one E&M service that was used to determine the episode's attribution to the TIN.

•Surgical episodes (i.e., episodes for which the index admission has a surgical MS-DRG) are attributed to the clinician/clinician group performing the main procedure during the inpatient stay. Specifically, the episode is attributed to the TIN and TIN-NPI who billed any related surgical procedure on Part B Physician/Supplier claims during the inpatient stay. The full list of Current Procedural Terminology / Healthcare Common Procedure Coding System (CPT/HCPCS) codes determined as related to each surgical MS-DRG can be found in the Measure Codes List file. See Section S.1. for link to Measure Codes List.

A few select surgical MS-DRGs are attributed using the 30% E&M rule, rather than a main procedure. During these surgical DRGs, the clinician(s) caring for the main disease process are the ones that are driving the care for the patient as opposed to the proceduralist who performed the primary procedure. For example, in an episode where a patient who has a prolonged intubation during an ICU stay and requires a tracheostomy, it would be unfair to attribute the proceduralist who performs the tracheostomy as they may only see the patient for this single procedure.

STEP 3: Exclude Clinically Unrelated Services to Calculate Episode Observed Cost

All Medicare Part A and Part B concurrent to the episode window are considered for inclusion toward the episode, with exceptions for services that are unlikely to be influenced by the clinician's care decisions. Services unlikely to be influenced by the clinician's care decisions are excluded based on rules developed by the MSPB Service Refinement Workgroup (discussed in Section S.8.3). The service exclusion rules are defined specific to the Major Diagnostic Category (MDC) of the index admission. The service exclusion codes and logic for services deemed clinically unrelated can be found in the "SE_[General/Post]_[Service_Category]" tabs of the Measure Codes List file (see Measures Codes List linked in Section S.1). After applying service exclusions, the standardized Medicare allowed amounts for the services included in each episode are summed to obtain the standardized episode observed cost.

Payment standardization is a process used by CMS to adjust the allowed charge for services to facilitate comparisons of resource use by removing geographic differences (e.g., due to labor costs) and adjustments from special Medicare programs (e.g., graduate medical education and disproportionate share payments). By

removing the effect of these factors, the payment standardization process preserves the differences in spending that are a result of healthcare delivery choices.

STEP 4: Exclude Episodes

A series of episode exclusions are applied to remove certain episodes from measure score calculation. Episodes are excluded from the MSPB Clinician measure if they meet any of the following conditions:

•Beneficiary has a primary payer other than Medicare during the episode window or in the 90-day lookback period

•Beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 90-day lookback period and episode window

•The beneficiary's death occurred during the episode.

•The index admission for the episode did not occur in either a subsection (d) hospital paid under the Inpatient Prospective Payment System (IPPS) or in an acute hospital in Maryland.

•The index admission for the episode is involved in an acute-to-acute hospital transfer (i.e., the admission ends in a hospital transfer or begins because of a hospital transfer).

•The index admission inpatient claim indicates a $0 actual payment or a $0 standardized payment.

After applying the exclusions outlined above, all remaining episodes are included in the calculation of the MSPB Clinician measure score.

Step 5: Calculate Expected Episode Costs Through Risk Adjustment

Risk adjustment is used to estimate episode expected costs in recognition of the different levels of care beneficiaries may require due to comorbidities, disability, age, and other risk factors. A separate risk adjustment model is estimated for episodes within each MDC, which is determined by the MS-DRG of the index admission. This model includes variables from the CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) 2016 Risk Adjustment Model and other standard risk adjustors to capture beneficiary characteristics.

Steps for defining risk adjustment variables and estimating the risk adjusted expected episode cost are as follows:

1) Define HCC and patient characteristic-related risk adjustors using Medicare Parts A and B claims in the 90-day lookback period from the episode start date.

2) Define other risk adjustors that rely upon Medicare beneficiary enrollment and assessment data as follows:

2a) Identify beneficiaries who are originally "Disabled without end-stage renal disease (ESRD)" or "Disabled with ESRD" using the original reason for joining Medicare field in the Medicare beneficiary enrollment database.

2b) Identify beneficiaries with ESRD if their enrollment indicates ESRD coverage, ESRD dialysis, or kidney transplant in the Medicare beneficiary enrollment database in the 90-day lookback period.

2c) Identify beneficiaries who are resident in a long-term care institution (90 days without having been discharged for 14 days) as of the episode start date using MDS assessment data.

3) Categorize beneficiaries into age ranges using their date of birth information in the Medicare beneficiary enrollment database.

4) Calculate an ordinary least squares (OLS) regression model to estimate the relationship between all the risk adjustment variables and the dependent variable, the standardized observed episode cost, to obtain the expected episode cost. A separate OLS regression is run for each episode MDC group nationally.

5) Winsorize the expected episode cost by assigning the value of expected episode cost at the 0.5th percentile of the distribution for episodes within the same MDC to all episodes with expected episode costs below the 0.5th percentile.

6) Renormalize values by multiplying each episode´s winsorized expected cost by the ratio of the MDC group´s average observed cost and the MDC group´s average winsorized expected cost.

7) Exclude episodes with outlier residuals to obtain finalized episodes with expected cost. This step is performed across all episodes regardless of the MDC group.

7a) Calculate each episode´s residual as the difference between the observed cost and the re-normalized, winsorized expected cost computed above.

7b) Exclude episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution.

7c) Renormalize all remaining episodes by multiplying their cost by the ratio of the average observed episode cost and the average winsorized expected cost when excluding outliers.

Step 6: Calculate Measure Scores

The MSPB Clinician measure is calculated for each clinician (TIN-NPI) or clinician group practice (TIN) by calculating the risk-adjusted episode cost ratio and multiplying the average cost ratio by the national average standardized episode cost. This method of cost ratio calculation allows for comparison of differences in observed and expected costs at the level of each individual episode before comparison at the clinician or clinician group level.

Specifically, the measure is calculated as follows:

•For each non-outlier episode, divide the episode's standardized observed cost by the episode's final expected cost to obtain the risk-adjusted episode cost ratio.

•Average the risk-adjusted cost ratios across all episodes for each TIN or TIN-NPI, and multiply this average cost ratio by the national average episode cost (all total standardized costs averaged over the universe of attributed, non-outlier episodes) to obtain the MSPB Clinician measure score for each TIN or TIN-NPI. Multiplying the ratio by the national average cost per episode is done to present the clinician's average cost measure score as a dollar amount rather than a ratio to be a more meaningful figure for clinicians.

S.7.2a. CONSTRUCTION LOGIC ATTACHMENT or URL: If needed, attach supplemental documentation (Save file as: S_7_2_Construction_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.

URL: See URL provided in Section S.1 for Measure Codes List

Please supply the username and password:

Attachment:

S.7.3. Concurrency of clinical events, measure redundancy or overlap, disease interactions *(Detail the method used for identifying concurrent clinical events, how to manage them, and provide the rationale for this methodology.)*

The MSPB Clinician measure includes Medicare Part A and Part B services that are furnished to a beneficiary during the episode. The MSPB Clinician measure avoids redundancy or overlap of clinical events by counting each service once within a given episode for the attributed clinician(s).

The MSPB Clinician measure allows an episode to overlap with other episodes. Consider for example, a patient who is admitted to and discharged from Hospital A, then admitted to Hospital B within 30 days. The first hospitalization would trigger an MSPB Clinician episode for the clinician(s) at Hospital A providing care for that beneficiary. The cost of the second hospitalization would be included as part of the first episode's costs. The second hospitalization would also trigger an MSPB Clinician episode for the clinician(s) at Hospital B who provide care for the beneficiary. The costs of the second hospitalization are included once in the episode for clinicians at Hospital A, and once in the episode for clinicians at Hospital B, so there is no double counting. Allowing for overlapping episodes is necessary to ensure continuous accountability between providers

throughout a beneficiary's trajectory of care, as clinicians at both hospitals share incentives to deliver high quality care at a lower cost to Medicare and engage in patient-focused care planning and coordination.

The measure accounts for disease interactions through its risk adjustment model based on the CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) 2016 model. In addition to the HCCs, the model includes disease interactions (e.g., Cancer * Immune Disorders). Further details about the risk adjustment model and disease interaction terms are included in Section S.8.2.

S.7.4. Complementary services *(Detail how complementary services have been linked to the measure and provide rationale for this methodology.)*

An MSPB Clinician episode excludes a defined list of Medicare Part A and Part B services that have been determined through expert input to be clinically unrelated to the care provided during an episode. The rationale for excluding these services is to ensure that the measure is assessing the overall costs for inpatient care, taking into consider the sphere of influence for attributed clinicians. A defined list of rules and Medicare claims codes is used to identify and exclude clinically unrelated services. This list is discussed in Section S.7.2, linked in Section S.1, and the rationale for developing this list is detailed in Sections S.8.1-S.8.3.

S.7.5. Clinical hierarchies *(Detail the hierarchy of codes or condition groups used and provide rationale for this methodology.)*

Clinical hierarchies are embedded in the risk adjustment model, described in Section S.7.2 and in more detail in Sections S.8.4 and S.8.5. The MSPB Clinician measure uses variables from CMS' Hierarchical Condition Category (HCC) model. This approach is adopted to ensure sufficient capture of the patient's comorbid disposition prior to the index hospital admission and allow more comprehensive risk adjustment of comorbid factors. The model suppress HCCs for less severe manifestations of a conditions when evidence for the more severe condition is found. This is done to prevent collinearity.

S.7.6. Missing Data *(Detail steps associated with missing data and provide rationale for this methodology (e.g., any statistical techniques to impute missing data)*

:

Since the MSPB Clinician measure uses claims data, we expect a high degree of data completeness.

CMS has in place several auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and to recoup any overpayments. CMS routinely conducts data analysis 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 payment. Specifically, CMS works with Zone Program Integrity Contractors (ZPICs), and formerly Program Safeguard Contractors (PSCs), to ensure program integrity; the agency also uses Recovery Audit Contractors (RACs) to identify and correct for underpayments and overpayments.

CMS also uses the Comprehensive Error Rate Testing (CERT) Program to ensure that Medicare payments are correct in accordance with coverage, coding, and billing rules. Between 2005 and 2017, CERT estimates that proper payment, which is payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments each year. The FY 2018 Medicare FFS program proper payment rate was 91.9 percent.[1] CMS continues to perform successful corrective actions and give providers additional education to ensure accurate billing.

To further ensure the completeness and accuracy of data for each beneficiary who opens an episode, the measure excludes episodes where beneficiary date of birth information (an input to the risk adjustment model) cannot be found in the EDB or the beneficiary death date occurs before the episode trigger date (an indication of errant data).

The MSPB Clinician measure also excludes episodes where the beneficiary is enrolled in Medicare Part C or has a primary payer other than Medicare in the 90-day lookback period and episode window. In such situations, Medicare Parts A and B claims data may not capture the complete clinical profile for the beneficiary needed to capture the clinical risk of the beneficiary in risk adjustment. Furthermore, Parts A and B claims data may not

capture all Medicare resource use if some portion of the beneficiary's care is covered under Medicare Part C. These steps ensure that we have complete claims data for beneficiaries included in the MSPB Clinician measure.

To ensure claims completeness and inclusion of any corrections, the measure was developed and calculated using data with a three month claims run-out from the end of the performance period.

[1] Comprehensive Error Rate Testing (CERT) Program. "Appendices Medicare Fee-for-Service 2018 Improper Payments Report". Table A6. https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicare-FFS-Compliance-Programs/CERT/Downloads/2018MedicareFFSSuplementalImproperPaymentData.pdf

S.7.7. Resource Use Service Categories (Units) (Select all categories that apply)

Inpatient services: Inpatient facility services

Inpatient services: Evaluation and management

Inpatient services: Procedures and surgeries

Inpatient services: Imaging and diagnostic

Inpatient services: Lab services

Inpatient services: Admissions/discharges

Other inpatient services

Ambulatory services: Outpatient facility services

Ambulatory services: Emergency Department

Ambulatory services: Pharmacy

Ambulatory services: Evaluation and management

Ambulatory services: Procedures and surgeries

Ambulatory services: Imaging and diagnostic

Ambulatory services: Lab services

Other ambulatory services

Durable Medical Equipment (DME)

Other services not listed

All Part A claims

All Part A and B claims

All Part A and B claims

S.7.8. Identification of Resource Use Service Categories (Units)

(For each of the resource use service categories selected above, provide the rationale for their selection and detail the method or algorithms to identify resource units, including codes, logic and definitions.)

The MSPB Clinician measure assesses the standardized allowed amounts of services by clinicians during an MSPB episode, which includes all Medicare Parts A and B claims that occur 3 days prior to the index admission through 30 days after the hospital discharge that are not excluded by the service exclusion rules. The logic conditions for the service exclusion rules are included in the "SE_[Service_Category]" tabs of the Measure Codes list file (linked in Section S.1). This identification approach allows the MSPB Clinician measure to capture the breadth of service categories that can be attributed to the clinician responsible for managing the beneficiary's care, while excluding services (and corresponding costs) that are unlikely to be influenced by the clinician's care decisions and that are considered clinically unrelated to the management of care.

S.7.8a. If needed, provide supplemental resource use service category specifications in either URL (preferred) or as an attachment (Save file as S.7.8a_RU_Service_Categories):

URL: See URL provided in Section S.1

Please supply the username and password:

Attachment:

Clinical Logic

S.8.1. Brief Description of Clinical Logic (Briefly describe your clinical logic approach including clinical topic area, whether or not your account for comorbid and interactions, clinical hierarchies, clinical severity levels and concurrency of clinical events.)

The measure aims to provide actionable information to clinicians providing care for beneficiaries during their hospital stay within the overall goal of enabling clinicians to provide cost-effective and high-quality care. The clinical logic is constructed to achieve this objective.

Clinical Topic Area: population-based measure for beneficiaries with acute inpatient hospitalizations

Comorbidity and Interactions: The risk adjustment model includes a series of interaction terms between comorbidities and applies a variant of the CMS-HCC risk adjustment model. The risk adjustment model is also used to account for clinical severity levels of beneficiary episodes.

Clinical Hierarchies: Clinical hierarchies are embedded in the risk adjustment model, based on the CMS-HCC model.

Additional clinical logic for this measure includes accounting for the attribution of episodes to clinicians using rules that reflect different types of clinical practice. The measure also excludes services that are clinically unrelated to the index admission through service exclusion rules defined by an expert clinician panel.

S.8.2. Clinical Logic *(Detail any clustering and the assignment of codes, including the grouping methodology, the assignment algorithm, and relevant codes for these methodologies.)*

To account for the clinical severity of patients, the measure is risk adjusted at the Major Diagnostic Category (MDC) level, using a combination of clinical indicators of CMS' Hierarchical Condition Category Version 22 (CMS-HCC V22) risk adjustment model (patient-level), indicators of the severity of the index hospitalization (patient-level, MS-DRG of index hospitalization and prior hospitalization indicator), indicators that rely on Medicare beneficiary enrollment and assessment data (patient level, e.g., ESRD coverage), and combinations thereof. The risk adjustment models are run within each MDC and with these indicators to support comparability across episodes. Further, the risk adjustment indicators are assessed over the 90 days preceding the episode to ensure that clinical events occurring near the episode window are captured and to minimize the loss of data for patients with a limited history of Medicare claims and administrative data. The indicators used for risk adjustment and the methodology are detailed in the Measure Information Form linked in Section S.1.

The MSPB Clinician measure attribution methodology differs by Medical MS-DRG and Surgical MS-DRG. Episodes that have a Medical MS-DRG are attributed to clinicians who bill at least 30 percent of E&M codes on Part B Physician/Supplier claims during the index hospitalization and episodes that have a Surgical MS-DRG are attributed to clinicians who bill any related surgical procedure on Part B Physician/Supplier claims during the index hospitalization. This approach is used to attribute the episode to clinicians who manage the beneficiary's care during the episode. The rules and codes used to attribute episodes to clinicians are listed in the workbook linked in Section S.1.

To identify services that are clinically unrelated to the index admission, the measure uses a defined list of service exclusion rules – these services are not included in the measure if they occur anytime during the episode window. This approach is used to limit the services and corresponding costs that clinicians are scored on to the services that are likely to be influenced by the clinician's care decisions and that are considered clinically related to the management of care. The defined list of service exclusion rules was developed with

input from clinician experts. The rules and codes used to attribute episodes to clinicians are listed in the Measure Codes List linked in Section S.1.

S.8.3. Evidence to Support Clinical Logic Described in S.8.2 *Describe the rationale, citing evidence to support the grouping of clinical conditions in the measurement population(s) and the intent of the measure (as described in IM3)*

The clinical logic used in the MSPB Clinician measure is informed by the literature, expert input, and feedback from a broad range of stakeholders.

The intent of the measure is to assess the resource use for patients receiving inpatient care. The rationale for assessing this area of care is in line with the overall goals of MIPS to evaluate costs, along with other domains such as quality, to reward clinicians who provide high-quality and cost-effective care. This measure is also intended to meet one of the Meaningful Measure areas and National Quality Strategy objectives to make care affordable. Cost measures such as the MSPB Clinician measure offer opportunity for improvement where clinicians can exercise influence on costs during the episode, or if lower spending and better care quality can be achieved through changes in clinical practice [1]. The inpatient setting is an area of high spending where increased cost effectiveness can be impactful in keeping Medicare spending affordable: in 2016, Medicare FFS paid $183 billion for approximately 10 million Medicare inpatient admissions and 200 million outpatient services, which reflects a 2.3 percent increase in hospital spending per FFS beneficiary between 2015 and 2016 [2]. Payment models like MIPS can have significant impacts on reducing costs and making care more affordable. The MSPB Clinician measure helps to assess the cost of care for patients to provide information to clinicians that can be used to curb health care costs.

Given that the inpatient hospital setting is such an important contributor to overall Medicare spending, it is necessary to measure costs related to hospitalizations. There is a strong association between high levels of total patient spending and the use of inpatient services [3,4,5]. Clinicians providing care in inpatient settings have opportunities to influence patient outcomes and cost performance, such as through reduced readmission rates.[6]

The clinical logic for the risk adjustment model is to appropriately account for patient comorbidities. Patient comorbidities are associated with higher resource use in the inpatient setting, such as through additional hospitalization charges, longer stays, and higher readmission rates. These include comorbidities for chronic conditions; for example, diabetes, hypertension, and heart failure have been found to be associated with higher levels of resource use [7,8]. Also, psychiatric comorbidities (e.g., depression, anxiety, dementia, substance use, bipolar disorders) have been associated with higher readmission rates for common inpatient treatment.[9,10] Medicare beneficiaries with multiple comorbidities account for a disproportionate amount of expenditure, including through additional resource use and length of stays [11,12]. As such, it is important to account for patient comorbidities and disease interactions in a resource use measure.

The clinical logic defines MSPB Clinician episode window as spanning 3 days prior to admission, through to 30 days post-discharge from an acute inpatient hospital. This episode definition is designed to align with the Medicare Spending Per Beneficiary – Hospital (NQF #2158) measure, which uses the same episode window. By aligning the episode window, the measure seeks to create aligned incentives for all clinicians and providers involved in providing care to a beneficiary who is admitted to hospital. This is also consistent with NQF's theoretical definition of an episode of care in that it is "…a series of temporally contiguous healthcare services related to the treatment of a given spell of illness or provided in response to a specific request by the patient or other relevant entity."[13]

The service exclusions list and attribution methodology were developed with input from a technical expert panel and a workgroup composed of clinicians with experience across different types of inpatient care. The measure was then field tested nationally to gather further input and feedback from the broader clinician community and other stakeholders. Further details about the development and testing process – including results of a vote to establish face validity - are contained in the Measure Testing Form, question 2b1.

[1] Fred, H L. “Cutting the Cost of Health Care: The Physician’s Role.” Texas Heart Institute Journal, vol. 43, no. 1, 2016.

[2] MedPAC. (2018) Report to the Congress: Medicare Payment Policy.”

[3] Lieberman SM, Lee J, Anderson T, et al. Reducing the growth of Medicare spending: geographic versus patient-based strategies. Health Aff 2003;Suppl Web Exclusives:W3-603–13

[4] Guo JJ, Ludke RL, Heaton PC, Moomaw CJ, Ho M, Cluxton RJ Jr Characteristics and risk factors associated with high-cost Medicaid recipients. Manag Care Interface. 2004 Oct; 17(10):20-7.

[5] Wammes, J J G et al. “Systematic review of high-cost patients´ characteristics and healthcare utilisation.” BMJ open vol. 8,9 e023113. 8 Sep. 2018, doi:10.1136/bmjopen-2018-023113

[6] Stevens, J ; Nyweide, D; Maresh, S, Comparison of Hospital Resource Use and Outcomes Among Hospitalists, Primary Care Physicians, and Other Generalists,. JAMA Intern Med. 2017;177(12):1781-1787. doi:10.1001/jamainternmed.2017.5824

[7] Boehme J, McKinley S, Michael Brunt L, Hunter TD, Jones DB, Scott DJ, Schwaitzberg SD.

Patient comorbidities increase postoperative resource utilization after laparoscopic and open cholecystectomy. Surg Endosc. 2016 Jun;30(6):2217-30. doi: 10.1007/s00464-015-4481-6. Epub 2015 Oct 1.

[8] Weeks, D L., Daratha KB, and Towle LA. “Diabetes Prevalence and Influence on Resource Use in Washington State Inpatient Rehabilitation Facilities, 2001 to 2007.” Archives of Physical Medicine and Rehabilitation 90, no. 11 (November 2009): 1937–43. https://doi.org/10.1016/j.apmr.2009.06.008.

[9] Sayers, SL., Hanrahan N, Kutney A, Clarke S, Reis BF, and Riegel B. “Psychiatric Comorbidity and Greater Hospitalization Risk, Longer Length of Stay, and Higher Hospitalization Costs in Older Adults with Heart Failure.” Journal of the American Geriatrics Society 55, no. 10 (October 2007): 1585–91. https://doi.org/10.1111/j.1532-5415.2007.01368.x

[10] Ahmedani, B. K., J. Hu, D. R. Nerenz, and L. K. Williams. “Psychiatric Comorbidity and 30-Day Readmissions after Hospitalization for Heart Failure, AMI, and Pneumonia.” American Psychiatric Association 66, no. 2 (February 1, 2015): 134–40

[11] Sorace, J, Millman M, Bounds M, Collier M, Wong H, Worrall C, Kelman J, and Macurdy T. “Temporal Variation in Patterns of Comorbidities in the Medicare Population.” Population Health Management 16, no. 2 (2013): 120–24. https://doi.org/10.1089/pop.2012.0045

[12] Pugely, A J., Martin C T, Gao Y, Belatti D A, and Callaghan J J. “Comorbidities in Patients Undergoing Total Knee Arthroplasty: Do They Influence Hospital Costs and Length of Stay?” Clinical Orthopaedics and Related Research® 472, no. 12 (May 2014): 3943–50. https://doi.org/10.1007/s11999-014-3918-x

[13] National Quality Forum. (2010). Measurement framework: Evaluating efficiency across patient-focused episodes of care. In Patient-Focused Episodes of Care. Retrieved from http://www.qualityforum.org/Publications/2010/01/Measurement_Framework__Evaluating_Efficiency_Across_Patient-Focused_Episodes_of_Care.aspx

S.8.3a. CLINICAL LOGIC ATTACHMENT or URL: If needed, attach supplemental documentation (Save file as: S_8_3a_Clinical_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.

URL: See URL provided in Section S.1

Please supply the username and password:

Attachment:

S.8.4. Measure Trigger and End mechanisms *(Detail the measure's trigger and end mechanisms and provide rationale for this methodology)*

Trigger Event: admission to acute care hospital (“index admission”)

MSPB Clinician Episode Start Date: 3 days prior to index inpatient hospital admission

MSPB Clinician Episode End Date: 30 days after discharge from the index inpatient hospital admission

The triggering and ending mechanism, in conjunction with the exclusionary clinical logic detailed in Sections S.8.2 and S.8.3, ensure the capture of a complete profile of services for the management of care surrounding an inpatient stay. The static timing of the episode start and end date are straightforward to ensure that clinicians can easily understand the episode window and construction, which is important for the measure's actionability, and is easily implementable since it includes all claims, except for a limited set of excluded services.

The 3 day prior to index admission period is motivated by Medicare's differential payment policies on services leading to an inpatient admission. Specifically, diagnostic services and non-diagnostic services that are related to the reason for inpatient admission and performed by the hospital are paid under the Inpatient Prospective Payment System (IPPS), while services furnished during this period are paid separately from the hospital payment if they are performed by a provider other than the hospital.

Services captured 30 days after a hospital discharge emphasize the importance of care transitions and care coordination. The length of this period was selected to align with other measures (e.g., NQF #2158 MSPB Hospital cost measure) which is long enough to capture costs related to the hospital stay, without being so long as to reduce the attributed clinicians' influence.

S.8.5. Clinical severity levels *(Detail the method used for assigning severity level and provide rationale for this methodology)*

Clinical severity levels are embedded in the risk adjustment methodology, which is based on the CMS-HCC model. That model, described in Section S.8.6, includes variables indicating a patient's health status at the start of the episode. In addition, the risk adjustment model adjusts for the MS-DRG of the index admission that triggered the episode, which reflects severity levels for that type of admission as there are separate MS-DRGs to indicate Complication and Comorbidity, Major Complication and Comorbidity, or no Complication and Comorbidity/Major Complication and Comorbidity.

In addition, the risk adjustment model includes status indicator variables for whether the beneficiary qualifies for Medicare through Disability or has ESRD. The model also includes an indicator of whether the beneficiary was receiving long-term care as of the start of the episode, defined as 90 days in a long-term care facility without being discharged to community for 14 days. Beneficiaries who need to reside in long-term care facilities typically require more intensive care than beneficiaries who live in the community. These enrollment and long-term care status variables are non-diagnostic based indicators of severity of illness.

S.8.6. Comorbid and interactions *(Detail the treatment of co-morbidities and disease interactions and provide rationale for this methodology.)*

Comorbidities and severity of illness are measured using HCCs, indicators of enrollment and long-term care status, and disease interactions. The risk adjustment model for the MSPB Clinician measure broadly follows the CMS-HCC risk adjustment methodology used in the Medicare Advantage (MA) program. The MSPB Clinician model includes 79 HCC indicators derived from the beneficiary's Parts A and B claims during the period 90 days prior to the episode start date, used in the CMS-HCC Version 22 (V22) 2016 model. The MSPB Clinician risk adjustment model includes 12 age categorical variables.

As the relationship between comorbidities' episode cost may be non-linear in some cases (i.e., beneficiaries may also have more than one disease during a hospitalization episode), the model also takes into account a limited set of interactions between HCCs and/or enrollment status variables. The risk adjustment methodology includes only a limited set of interaction terms for two reasons. First, inclusion of too many interaction terms will over-fit the model. Second, the risk-adjustment methodology broadly follows the established CMS-HCC risk-adjustment methodology, which uses similar interaction terms.

Adjustments for Comparability

S.9.1. Inclusion and Exclusion Criteria *Detail initial inclusion/exclusion criteria and data preparation steps (related to clinical exclusions, claim-line or other data quality, data validation, e.g. truncation or removal of low or high dollar claim, exclusion of ESRD patients)*

:

Included population:

The beneficiary population eligible for the MSPB Clinician measure calculation consists of Medicare beneficiaries enrolled in Medicare Parts A and B who had an index admission to an inpatient hospital. To be included, the beneficiary must have an episode ending during the performance period.

Exclusions:

Several steps in the construction of the MSPB Clinician measure ensure comparability of the MSPB Clinician measure by fostering comparability in the service profiles and population captured by the measure, as discussed in Section S.7.2.

The measure excludes services that are clinically unrelated to clinician care management or the index hospitalization furthers the comparability of services captured by measure by limiting service variation to services that are likely to be influenced by clinician care management and related to the index admission. This is Step 3 of the measure construction methodology.

The measure excludes select episodes, detailed in Step 4 of the measure construction methodology, furthers the comparability of the Medicare beneficiary population studied by excluding episodes if any of the following conditions are met:

•Beneficiary has a primary payer other than Medicare during the episode window or in the 90-day lookback period

•Beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 90-day lookback period and episode window

•The beneficiary's death occurred during the episode.

•The index admission for the episode did not occur in either a subsection (d) hospital paid under the Inpatient Prospective Payment System (IPPS) or in an acute hospital in Maryland.

•The index admission for the episode is involved in an acute-to-acute hospital transfer (i.e., the admission ends in a hospital transfer or begins because of a hospital transfer).

•The index admission inpatient claim indicates a $0 actual payment or a $0 standardized payment.

The rationale and testing results for these exclusions are contained in the testing attachment, Section 2b2.

The MSPB Clinician measure applies risk adjustment, statistical exclusions, and renormalization to further ensure comparability, described in Step 5 of the construction methodology. The risk adjustment approach accounts for patient level variation prior to the index hospitalization and the severity of the index hospitalization. Statistical exclusions and renormalizations are engaged during measure construction after excluding outlier episodes to ensure that distributions resulting from outlier exclusions remain true to population averages.

As with the CMS-HCC model, the risk adjustment approach for this measure uses an ordinary least squares linear regression model. The predicted, or expected, cost is winsorized at 0.5th percentile to make sure episodes with unusually small predicted cost, which would lead to abnormally large O/E ratios, do not dominate certain clinicians' final score. The winsorized expected costs are renormalized to ensure the average expected episode cost is the same before and after winsorizing. Then, extremely low- or high-cost outlier episodes with residuals below the 1st percentile or above the 99th percentile are excluded to reduce the effect of these episodes that deviate the most from their expected values in absolute terms. The expected cost after excluding these outliers is again renormalized to ensure that average expected costs are the same after outlier removal.

S.9.2. Risk Adjustment Type (Select type)

Stratification by risk category/subgroup

If other:

S.9.3. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets)*

The MSPB Clinician measure is stratified by MDC, which are mutually exclusive groups of MS-DRGs that correspond to an organ system (e.g., diseases and disorders of the digestive system) or cause of admission (e.g., burns). There are 25 MDCs (numbered 01-25), and a Pre-MDC group for extremely resource intensive MS-DRGs. Unlike MS-DRGs within the numbered MDCs which are determined largely by principal diagnosis, MS-DRGs within the Pre-MDC group are determined by Operating Room procedures (e.g., organ transplant). By running the risk adjustment model described in Section S.7.2 separately for episodes within each MDC determined by the MS-DRG of the index admission, the MSPB Clinician measure accounts for differences in resource use due to the nature of the reason for hospitalization. This helps ensure that the cost measure is fairly comparing clinicians for their patient case-mix, while preserving clinically meaningful distinctions in the beneficiary population within each MDC.

The risk adjustment variables included in the model are listed in the Measure Codes List linked in Section S.1.

S.9.4 Costing method

Detail the costing method including the source of cost information, steps to capture, apply or estimate cost information, and provide rationale for this methodology.

Standardized pricing

The measure removes sources of variation in spending that are unrelated to healthcare delivery choices, as described in Section S.7.2. The methodology used to payment standardize the Medicare claims used to specify this measure is available for download ("CMS Price (Payment) Standardization") from the following URL: https://www.qualitynet.org/inpatient/measures/payment-standardization

S.10. Type of score*(Select the most relevant):*

Ratio

If other:

Attachment:

S.11. Interpretation of Score *(Classifies interpretation of a ratio score(s) according to whether higher or lower resource use amounts is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score, etc.)*

The MSPB Clinician measure score is presented as a dollar value that represents the average payment-standardized, risk-adjusted cost across all MSPB Clinician episodes attributed to a given clinician/clinician group. A lower measure score indicates that the observed episode costs are lower than or similar to expected costs for the care provided for the particular patients and episodes included in the calculation. A higher measure score indicates that the observed episode costs are higher than expected for the care provided for the particular patients and episodes included in the calculation.

S.12. Detail Score Estimation *(Detail steps to estimate measure score.)*

As described in Section S.7.2, the MSPB clinician measure is calculated for each clinician (TIN-NPI) or clinician group practice (TIN) by (i) calculating the ratio of standardized observed episode costs to final expected episode costs and (ii) multiplying the average cost ratio across episodes for each TIN or TIN-NPI by the national average standardized episode cost. This method of cost ratio calculation allows for comparison of differences in observed and expected costs at the level of each individual episode before comparison at the clinician or clinician group level.

1) Calculate risk-adjusted episode cost ratio. For each non-outlier episode, the episode's total standardized observed cost is divided by the episode's final expected cost.

2) Calculate the MSPB Clinician measure for each TIN or TIN-NPI. After calculating each episode's risk-adjusted cost ratio, average this cost ratios across all episodes for each TIN or TIN-NPI. Multiplying this average cost ratio by the national average episode cost (all total standardized costs averaged over the universe of attributed, non-outlier episodes) gives the MSPB clinician measure for each TIN or TIN-NPI. Multiplication of the ratio by national average cost per episode is done to convert the ratio into a figure that is more meaningful from a cost perspective by having the clinician's average cost measure score represented as a dollar amount rather than a ratio.

Reporting Guidelines

This section is optional and will be available for users of the measure as guidance for implementation and reporting.

S.13.1. Describe discriminating results approach

Detail methods for discriminating differences (reporting with descriptive statistics--e.g., distribution, confidence intervals).

This version of the MSPB Clinician measure that underwent comprehensive re-evaluation in 2018 and rulemaking in 2019 will be reported as part of the MIPS Cost Performance Category for the CY 2020 performance period onwards. The Cost Performance Category score is calculated as the equally weighted average of all cost measures for which a clinician has the required number of cases. The Cost Performance Category score will make up 15% of the composite MIPS Final Score in CY 2020, balanced with scores from the other performance categories: Quality (45%), Improvement Activities (15%), and Promoting Interoperability (25%). While this measure does capture consequences of care such as complications, there are other quality metrics that cannot be captured by a cost measure alone. As such, this measure is most meaningful when reported as part of a program such as MIPS where clinicians are also assessed on quality measures.

While this version of the MSPB Clinician measure has not yet been reported as part of MIPS, the clinician community has had opportunities to review and become familiar with the revised measure. During measure development, we conducted national field testing in October 2018 where a total of nearly 150,000 field test reports containing cost measure performance on the draft MSPB Clinician measure as specified at that time were available to clinicians and clinician groups meeting a 35-episode case minimum (20,852 TIN-level reports and 127,530 TIN-NPI level reports). During field testing, a National Summary Data Report was also posted containing summary statistics, including information on the distribution of TIN and TIN-NPI level measure scores.

S.13.2. Detail attribution approach

Detail the attribution rules used for attributing resources/costs to providers (e.g., a proportion of total measure cost or frequency of visits during the measure's measurement period) and provide rationale for this methodology.

As described in Step 2 in Section S.7.2, the MSPB Clinician measure uses a separate attribution methodology for medical and surgical episodes. An episode with a medical MS-DRG is attributed to a TIN if that TIN billed at least 30 percent of the evaluation and management (E&M) claims billed during the inpatient stay, and to a TIN-NPI if the clinician within an attributed TIN billed at least one E&M claim that was used to determine the episode's attribution to the TIN. The list of inpatient E&M codes used for this attribution can be found in the Measure Codes List File linked in Section S.1 on the "Med_Attribution_E&M" tab.

An episode with a surgical MS-DRG is attributed to a TIN if that TIN billed the relevant CPT/HCPCS code determined to be related to the surgical MS-DRG, and to a TIN-NPI if the clinician billed the relevant CPT/HCPCS code determined to be related to the surgical MS-DRG. The list of relevant CPT/HCPCS codes used to attribute surgical episodes can be found in the Measure Codes List File on the "Surg_Attribution_CPT_HCPCS" tab.

A few select surgical MS-DRGs are attributed using the 30% E&M rule, rather than a main procedure. During these surgical DRGs, the clinician(s) caring for the main disease process are the ones that are driving the care for the patient as opposed to the proceduralist who performed the primary procedure. The list of MS-DRGs and their respective attribution rule can be found in the Measure Codes List File on the "Attribution_Rule" tab.

Based on input from a technical expert panel, this attribution methodology was incorporated as a refinement as part of the comprehensive re-evaluation process of the MSPB measure as specified in MIPS for performance years 2017 through 2019. This revised attribution methodology accounts for the team-based nature of care provided when managing medical conditions during an inpatient stay and allows for attribution to multiple clinicians to ensure that all clinicians involved in a beneficiary´s care are appropriately attributed.

S.13.3. Identify and define peer group

Identify the peer group and detail how peer group is identified and provide rationale for this methodology.

The peer group for this measure includes all clinicians and clinician groups providing care to beneficiaries in an inpatient setting, as identified by meeting the triggering and attribution logic. The rationale for identifying the peer group as clinicians providing care in inpatient hospitals is to assess the resource use of clinicians for the cost performance category of MIPS; this program and the requirement to have a cost performance category was established by the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). The measure ensures clinical comparability through the techniques described in Sections S.9.1-S.9.4.

S.13.4. Sample size

Detail the sample size requirements for reporting measure results.

From the MIPS CY 2020 performance period and onwards, the MSPB Clinician measure will be calculated and reported via confidential reports for TINs and TIN-NPIS with 35 or more episodes. Public reporting may be introduced for MIPS cost measures in the future.

S.13.5. Define benchmarking and comparative estimates

Detail steps to produce benchmarking and comparative estimates and provide rationale for this methodology.

The measure is not calculated against a benchmark, but as a ratio of the attributed clinician's observed over expected costs across their MSPB Clinician episodes, multiplied by the national observed episode cost to generate a dollar figure. It will be used in the MIPS Cost Performance Category for the 2020 performance period onwards. Reporting this measure as part of the cost performance category helps to measure clinicians' resource use for services they administer to Medicare beneficiaries related to inpatient hospitalizations to hold clinicians accountable for their cost effectiveness. Combined with measures in the other MIPS performance categories, such as the quality performance category, the MSPB Clinician measure allows CMS to assess the value of care and incentivize both achievement and improvement in the provision of high-quality, cost-effective care.

Validity – See attached Measure Testing Submission Form

SA.1. Attach measure testing form

2020-04-29-nqf-testing-form-mspb-clinician-v8.docx

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** N/A
Measure Title: Medicare Spending Per Beneficiary (MSPB) Clinician
Date of Submission: 4/29/2020

Type of Measure:

☐ **Outcome (*including PRO-PM*)**	☐ **Composite – *STOP – use composite testing form***

☐ Intermediate Clinical Outcome	☒ Cost/resource
☐ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	

Instructions

- **Measures must be tested for all the data sources and levels of analyses that are specified. *If there is more than one set of data specifications or more than one level of analysis, contact NQF staff* about how to present all the testing information in one form.**
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures, section 2b3 also must be completed.**
- **If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), section 2b5 also must be completed.**
- **Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.**
- **If you are unable to check a box, please highlight or shade the box for your response.**
- **Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). *Contact NQF staff if more pages are needed.***
- **Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.**
- **For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.**

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care

(e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment	☒ other: Long-term Minimum Data Set (assessment data), Enrollment Database, Common Medicare Environment, American Community Survey (ACS)

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

The MSPB Clinician measure uses Medicare Part A and Part B claims data maintained by CMS. Part A and B claims data are used to build episodes of care, calculate episode costs, and construct risk adjustors. Data from the EDB is used to determine beneficiary-level exclusions and supplemental risk adjustors, specifically Medicare Parts A, B, and C enrollment, primary payer, disability status, end-stage renal disease (ESRD), beneficiary birth dates, and beneficiary death dates. The risk adjustment model also accounts for expected differences in payment for services provided to beneficiaries in long-term care based on the data from the MDS. Specifically, the M DS is used to create the long-term care indicator variable in risk adjustment.

For measure testing, data from the American Census, American Community Survey (ACS), and CME are used in analyses evaluating patient cohort and social risk factors in risk adjustment.

1.3. What are the dates of the data used in testing? MSPB Clinician episodes ending from January 1, 2018 to December 31, 2018. The split-sample intraclass correlation also includes episodes ending in the 2017 calendar year. For further details, please see Question 1.7.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.20*)	Measure Tested at Level of:
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

There were 19,213 clinician group practices (identified by Tax Identification Number [TIN]) and 126,628 practitioners (identified by combination of TIN and National Provider Identifier [NPI]) included in the analysis. Clinicians and clinician groups were included if they were attributed 35 or more MSPB Clinician episodes, as identified in Medicare Parts A and B claims data, ending from January 1, 2018, to December 31, 2018. Episodes from all 50 States and D.C. with an index admission in the acute inpatient setting were included.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

There were 4,114,268 Medicare beneficiaries (from 6,028,412 episodes) included in TIN level measure testing, and 3,755,580 beneficiaries (from 5,442,789 episodes) included in TIN-NPI level measure testing. The beneficiaries included in the MSPB Clinician measure calculation are enrolled in Medicare Parts A and B (but not Part C) and have had an admission to an acute care hospital. Beneficiaries and their episodes were included in the sample if they met a set of inclusion criteria (listed below) meant to ensure completeness of data and to focus the measure on a clinically homogeneous cohort of patients. The inclusion criteria are:

- The beneficiary has Medicare as their primary payer for the entire time during the episode window and 90-day lookback period prior to the episode start day used for risk adjustment.
- The beneficiary was continuously enrolled in Medicare Parts A and B for the entirety of the lookback period plus episode window, and was not enrolled in Medicare Part C for any time during this duration.
- The index admission of the episode was in an acute inpatient facility located in the United States.
- The beneficiary date of birth is not missing.
- The beneficiary death date did not occur before the episode end date.
- The index admission for the episode occurred in either a subsection (d) hospital paid under the Inpatient Prospective Payment System (IPPS) or in an acute hospital in Maryland[1]
- The index admission for the episode was not involved in an acute-to-acute hospital transfer (i.e. the admission does not end in a hospital transfer or does not begin because of a hospital transfer)
- The claim for the index admission indicated a positive actual and standardized payment.

[1] Subsection (d), which covers hospitals in the 50 states and D.C., does not include psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer.

To determine whether the MSPB Clinician measure's inclusion criteria distort patient characteristics on episodes, we produced and analyzed distributions of patient characteristics (age, race, sex, dual eligibility status, income, unemployment, hierarchical condition categories [HCCs]) for (i) episodes with inclusion criteria, (ii) episodes without inclusion criteria, (iii) beneficiaries with inclusion criteria, and (iv) beneficiaries without inclusion criteria.

This analysis shows that the MSPB Clinician measure's inclusion criteria have only a minimal effect on the percentage of beneficiaries of any particular demographic (Appendix Table 1.6). The difference between the proportion of beneficiaries observed for each of the demographic before and after applying inclusion criteria is between -1.7 and 1.7 percentage points. To illustrate, the percentage of beneficiaries aged 65 to 69 without applying the inclusion criteria is 17.8 percent while the percentage of beneficiaries aged 65 to 69 with applying the inclusion criteria is 16.7 percent. The breakdown of male and female beneficiaries is similar when comparing the use of inclusion criteria, with 54.1 percent without inclusion criteria and 55.7 percent with inclusion criteria for female. These results indicate that there is minimal shift in patient characteristics after application of the inclusion criteria listed above.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

The calculation of a split-sample intraclass correlation to test reliability (Section 2a2) aggregates episodes ending in calendar years 2017 and 2018. All other testing used the study period of January 1, 2018 to December 31, 2018.

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

The social risk factors analyzed were variables from the ACS, EDB, and CME. All ACS variables firstly defined at the Census Block Group level and then ZIP code when census block group is missing. Social risk variables analyzed include the following:

- Income (ACS): Low Income: median income < 33rd percentile nationally; Medium Income: median income in the interval spanning the 33rd percentile to the 66th percentile nationally; High Income: median income > 66th percentile
- Education (ACS): Education < High School: when % with < high school education is the highest for a given Census Block Group; Education = High School: when % with only high school is the highest; Education > High School: when % with > high school is the highest
- Employment (ACS): Unemployment Rate > 10%; Unemployment Rate <= 10%
- Race (EDB): Asian, Black, Hispanic, North American Native, White, and Other
- Sex (EDB): Female, male
- Dual status (CME): Full dual, partial dual, non-dual
- Agency of Healthcare Research and Quality (AHRQ) SES Index: AHRQ index scores are calculated using the AHRQ scoring algorithm and is a continous dependent variable as a replacement of all SES variables. The index includes percentage of households containing one or more person per room, median value of owner-occupied dwelling, percentage of persons below the federally defined poverty line, median household income, percentage of persons aged ≥ 25 years with at least 4 years of college,

percentage of persons aged ≥ 25 years with less than a 12th grade education, and percentage of persons aged 16 or older in the labor force who are unemployed.[2] [3]

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability Score: 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 and σ_b^2 is the between-group variance of clinician 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.

Split Sample Reliability Testing: This test examined agreement between two performance measure scores for a TIN or TIN-NPI based on random-split, independent subsets of episodes from an aggregation of two years of episodes (2017-2018). We used two years of data to achieve numbers of episodes per TIN or TIN-NPI that are comparable to the number of episodes in one year, as this measure is calculated and reported for a one-year

---

[2] Agency for Healthcare Research & Quality, Centers for Medicare & Medicaid Services, and RTI International. "Creation of New Race-Ethnicity Codes and Socioeconomic Status (SES) Indicators for Medicare Beneficiaries." Research Triangle Park, 2008. https://archive.ahrq.gov/research/findings/final-reports/medicareindicators/index.html

[3] SES Index Score = 50 + (-0.07 * [% of households containing one or more person per room]) + (0.08 * [median value of owner-occupied dwelling, standardized range from 0-100] + (-.010 * [% of persons below the federally defined poverty line]) + (0.11 * [median household income, standardized range from 0-100]) + (0.10 * [% of persons aged ≥ 25 years with at least 4 years of college] + (-0.11 * [% of persons aged ≥ 25 years with less than a 12th grade education]) + (-0.08 * [% of persons aged 16 or older in the labor force who are unemployed])

performance period in MIPS. Good agreement indicates that the performance score is more the result of a TIN or TIN-NPI group's characteristics, like efficiency of care, rather than statistical noise due to random variation. Only TIN and TIN-NPIs that met a case minimum of 35 episodes in both performance periods were included. The sample was stratified by the calendar years, thus ensuring that episodes within each calendar year were evenly distributed across the split-halves for a given TIN or TIN-NPI. The split-half samples were used to calculate each sample's performance measure scores using the same specification. We then calculated Shrout-Fleiss intraclass correlation coefficients ICC(2,1) between the different performance scores to measure reliability. Lower ICC scores indicate less correlation between the two estimates, a score of 1 would mean the estimates are exactly the same.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Reliability Score Results. Table 1 presents the distribution of reliability scores for TINs and TIN-NPIs overall. At a testing volume threshold of at least 35 episodes (the case minimum for the measure in the MIPS 2020 performance period) the mean reliability for TINs is 0.78 and for TIN-NPIs is 0.70. 100 percent of TINs and TIN-NPIs at the reporting case minimum have reliability greater than or equal to 0.4, the standard that CMS generally considers as the threshold for 'moderate' reliability[4]. Mean reliability increases with increasing volume thresholds. While higher volume thresholds yield even higher reliability results, it is at the cost of further reducing the number of clinicians and clinician groups able to receive a measure score.

Table 1. Distribution of Reliability Scores for TINs and TIN-NPIs with an Overall Testing Volume Threshold of 35 Episodes

Reporting Level	Number of TINs or TIN-NPIs	Mean (Std. Dev.)	25th Pct.	50th Pct.	75th Pct.
TIN	19,213	0.78 (0.13)	0.67	0.79	0.90
TIN-NPI	126,628	0.70 (0.11)	0.60	0.69	0.79

* Pct. = percentile.

In response to stakeholder interest in seeing the measure reliability for clinician groups of different practice size, **Table 2** shows the distribution of reliability scores by the number of TIN-NPIs within a TIN. When examined by number of clinicians within the practice, the average reliability scores increases from 0.70 (1 clinician) to 0.90 (21+ clinicians) for TINs.

Table 2. Distribution of Reliability Scores for TINs by Practice Size, with an Overall Testing Volume Threshold of 35 Episodes

# of Clinicians	Number of TINs or TIN-NPIs	Mean (Std. Dev.)	25th Pct.	50th Pct.	75th Pct.
Overall	**19,213**	**0.78 (0.13)**	**0.67**	**0.79**	**0.90**
1 Clinician	5,771	0.70 (0.10)	0.62	0.69	0.77

[4] 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.

2-4 Clinicians	4,022	0.74 (0.10)	0.66	0.74	0.83
5-20 Clinicians	4,739	0.81 (0.11)	0.73	0.83	0.90
21+ Clinicians	4,681	0.90 (0.11)	0.85	0.94	0.98

* Pct. = percentile.

Split-sample Reliability Testing Results. Table 3 presents ICC(2,1) between the split-sample measures scores for the overall sample of 17,427 TINs and 95,647 TIN-NPIs included in this testing. The ICC in the overall sample for TIN reporting was 0.66 and 0.60 for TIN-NPI reporting.

Table 3. Split-sample Intraclass Correlation Coefficients

Reporting Level	# of TINs or TIN-NPIs	Mean Score: Sample 1	Mean Score: Sample 2	Pearson Correlation Coefficient	ICC(2,1)
TIN	17,427	1.0132	1.0132	0.66	0.66
TIN-NPI	95,647	1.0412	1.0413	0.60	0.60

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Overall, testing results indicated good or high measure score reliability with an average of 0.78 for TINs and 0.70 for TIN-NPIs at a volume threshold of 35 episodes.[5] Reliability for groups of different practice sizes was similar, with mean reliability for the smallest TINs at 0.70.

The split-sample reliability analysis provides further evidence of reliability and repeatability of the performance measure. Reliability (ICC(2,1)) was 0.66 for TINs and 0.60 for TIN-NPIs, which indicates substantial or moderate overall reliability for TINs and moderate for TIN-NPIs.

The two reliability metrics capture related, but distinct, concepts. Our ICC(2,1) metric will tend to differ from our signal-to-noise metric for two reasons: (i) The denominator of ICC(2,1) includes additional statistical variation arising from true differences in a provider's performance across performance periods; and (ii) The denominator of ICC(2,1) imposes a common variance for the residual across providers, ignoring differences in

[5] Thresholds for sufficient measure reliability (including the ICC and other reliability methods) vary across sources (see, for example, Portney and Watkins, 2000, for a discussion). Authors provide a range of thresholds; for example, Landis and Koch (1977) classify Kappa statistics in the 0.41-0.60 range as "moderate," 0.61-0.80 range as "substantial," and 0.81-1.00 range as "almost perfect." Koo and Li (2016), on the other hand, classify ICC values in the 0.5-0.75 range as "moderate," 0.75-0.9 range as "good," and above 0.9 as "excellent." Nunnally (1978) is often cited to justify a threshold of 0.7 for "sufficient" reliability. CMS provides the following thresholds: "*We generally consider reliability levels between 0.4 and 0.7 to indicate "moderate" reliability and levels above 0.7 to indicate "high" reliability.*" (Quality Payment Program 2017 Final Rule: 81 FR 77169). The Department of Education provides the following thresholds: "*Reliability of an outcome measure may be established by meeting the following minimum standards: (a) internal consistency (such as Cronbach's alpha) of 0.50 or higher; (b) temporal stability/test-retest reliability of 0.40 or higher; or (c) inter-rater reliability (such as percentage agreement, correlation, or kappa) of 0.50 or higher.*" (What Works Clearinghouse (WWC) Standards Handbook v4, p.78).

precision arising from differences in case sizes. Reason (i) makes ICC(2,1) a less relevant metric in this context, since program goals actually require accurately distinguishing systematic performance changes from one period to another, rather than treating them as statistical noise. To avoid this issue, one could alternatively calculate ICC(2,1) using split-half samples from a single performance period. However, this approach also underestimates reliability of the measure for use in the program; in this case, under-estimation occurs because case sizes are artificially cut in half from true case sizes, mechanically reducing precision from the intended application of the measures. We still present both reliability metrics for completeness, but for reasons (i) and (ii), view the signal-to-noise metric as the preferred and more relevant one.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Face validity

The MSPB Clinician measure underwent a structured process for gathering detailed input from recognized clinician experts on inpatient care. These expert panels were convened to provide input to inform the comprehensive measure re-evaluation process. Experts provided input on measure refinements (e.g., attribution logic and service exclusion rules) that would help ensure that the measure is fulfilling its intent to capture the overall costs of care for patients who had an inpatient stay (i.e., capturing what it is intending to capture and differentiating between provider performance).

During measure re-evaluation, we incorporated input from (i) a technical expert panel (TEP) which convened to discuss this measure at three meetings in 2017 and 2018, (ii) the MSPB Service Refinement workgroup, which convened in the summer of 2018, and (iii) stakeholder feedback from national field testing.

The TEP comprised 19 members with expertise in cost measure development and evaluation and quality improvement from diverse backgrounds, including clinicians, healthcare providers, academia, and patient advocacy organizations. The MSPB Service Refinement workgroup was composed of 25 members (including four TEP members) with experience and expertise in a broad range of inpatient care, affiliated with 21 specialty societies, including the American Medical Association, Society of General Internal Medicine, Society of Thoracic Surgeons, and the American Academy of Family Physicians.

At a TEP meeting in August 2017, the panel provided high-level guidance and initial input on direction of refinements, focusing on attribution and service refinements which stakeholders previously commented on for the version of the measure as specified at that time. These refinements would address prior feedback and help the measure ensure it was capturing the clinicians responsible for providing inpatient care, which is key to measure actionability and meaningfully differentiating between providers. During a May 2018 meeting, the TEP provided further input on specific approaches with empirically testing to refining attribution methodology and creating services exclusion logic.

The TEP also recommended the creation of a targeted MSPB Service Refinement workgroup to provide detailed clinical input on service assignment rules. During two webinars in June and July 2018, the MSPB Service Refinement workgroup reviewed and discussed empirical analyses and used their clinical expertise to provide input on developing service assignment exclusions for the measure, with exclusions specific to Major Diagnostic Category (MDC) groupings. This approach allowed for exclusions specific to organ system or reason for admission which members believed to be an appropriate level of granularity for the types of services to exclude as being clinically unrelated to the inpatient stay in consideration of the clinician's sphere of influence. Input was gathered in a structured manner including the use of a polling process requiring greater than 60 percent consensus.

In addition, a national field testing feedback period in October and November 2018 offered all stakeholders an opportunity to review and provide input on draft measure specifications for the refinements and measure feedback reports for attributed clinicians and clinician groups. During this period, 148,382 field test reports (20,852 for TINs and 127,530 for TIN-NPIs) were available for download and review for the MSPB Clinician measure revised in 2018. During a November 2018 meeting, the TEP reviewed feedback received on the measure from field testing and confirmed that they did not believe further refinements needed to be made to the measure.

To gather a formal record of the TEP's systematic input and iterative assessments of the measure refinements throughout this process, TEP members completed a face validity survey in November 2019 that assessed (i) the revised measure as compared to the previous version, and (ii) the measure as currently specified after refinements were made. The survey used a Likert scale with values of 1 = Strongly Disagree, 2 = Moderately Disagree, 3 = Somewhat Disagree, 4 = Somewhat Agree, 5 = Moderately Agree, and 6 = Strongly Agree. Fifteen of the 19 TEP members completed the survey.

Empirical Validity Testing

We undertook two approaches to empirically examine the extent to which the measure captures what it intends to capture. In the first approach, we sought to confirm the expectation that the MSPB Clinician measure captures variation in service utilization by examining differences in risk-adjusted cost for known indicators of resource or service utilization, specifically with downstream acute readmission and with post-acute care (PAC) service utilization. For this analysis, we compared the ratio of observed to expected cost (henceforth called the "O/E cost ratio") for MSPB Clinician episodes with and without readmissions and with or without PAC services utilization. We expected that episodes with these downstream services would be more expensive than episodes without.

In the second approach, we empirically tested whether the measure is capturing variation in provider cost in the manner intended by evaluating how different types of cost impact risk-adjusted measure scores. Certain services or costs included in the MSPB Clinician measure were classified into clinically coherent groups of services, called "clinical themes", and are:

- **Acute Inpatient Service**, including acute inpatient hospital index admission, and services billed by any clinician during index hospitalization
- **Inpatient Readmissions,** including acute inpatient hospitalization following the index admission and the related services billed by any clinician
- **Post-Acute Care (PAC)**, including home health (HH), skilled nursing facility (SNF), and inpatient rehabilitation or long-term care facility (IRF/LTCH)
- **Emergency Services Not Included in a Hospital Admission,** including emergency E&M services; procedures; laboratory, pathology, and other tests; and imaging services.

- **Outpatient Evaluation and Management Services, Procedure, and Therapy (excluding emergency department)**, including physical, occupational, or speech and language pathology therapy; E&M services, major procedures; anesthesia, and ambulatory/minor procedures.

We calculated the Pearson correlation between the cost of each clinical theme during the episode and the overall risk-adjusted cost for an episode. Also included are correlations between the cost of each clinical theme and the expected episode spending as predicted by risk-adjustment to show if the resource use within a clinical theme would be expected due to the patient pre-existing conditions outside the influence of the clinician. We hypothesized that the readmission service category would have the highest correlation with risk-adjusted episode cost, as readmissions are likely associated with high cost even after accounting for beneficiary characteristics. Similarly, we expected that the PAC: SNF service category would have a high correlation with risk-adjusted episode cost as well, since SNF services are frequently provided to beneficiaries after qualifying hospitalizations and tend to be less cost efficient than other PAC services (e.g., home health).

We also examined the possibility of testing a hypothesized relationship between clinicians' MSPB scores and their scores on MIPS quality measures. This type of testing assesses whether clinicians with better MSPB Clinician scores also perform well on quality measures aimed at capturing the same dimension of care. However, any relationship between cost measures and quality measures depends on many factors, including the exact construction of each measure, such as whether the specifications are sufficiently harmonized to be capturing the same patient care experience. It also depends on the dimension of care that the quality measure is assessing; for instance, outcomes measures may not be able to assess patient health indicators if the relevant outcome is unlikely to emerge in the short or medium term. An additional consideration specific to MIPS is the availability of quality measure data given reporting requirements. Participants in MIPS select six quality measures to report out of a large number of measures; in 2017 and 2018, there were over 270 available quality measures for clinicians to select from.

As there must be a conceptual basis for testing whether an expected relationship between cost and quality measures exists, stronger potential quality measures should capture the same dimension of care as the cost measure. For example, a quality measure that assesses the incidence of complications for the same patient cohort should conceptually be reflected in a cost measure that includes the cost of those complications, although there may be other factors driving costs besides those assessed by the quality measure. This kind of analysis looks into a hypothesized relationship between how different types of measures can test the claims-based MSPB Clinician measure against one that uses a different data source (e.g., health records). As such, we focused on identifying potential non-claims based outcomes measures for internal medicine and general surgery specialties that would be reflected in the cost measure. Of the available outcome measures, some were aimed at different types of care from what MSPB Clinician aims to capture (e.g., MIPS Q#342 is specific to admission to palliative care services) and others were very narrowly specified relative to the intent of MSPB Clinician (e.g., MIPS Q#356 Anastomotic Leak Intervention is only for gastric bypass and colectomy surgeries). Two eCQM measures - MIPS Q#355 Unplanned Reoperation within the 30 Day Postoperative Period, MIPS Q#356 Unplanned Hospital Readmission within 30 Days of Principal Procedure - have more potential to show a relationship with MSPB as these adverse outcomes for surgery and hospital admissions will also be captured by the MSPB Clinician measure in higher resource use. However, only 1,411 and 826 MIPS participants (clinicians and groups) reported Q#355, and Q#356, respectively, in 2017.[6] Even if there was complete reporting data, the patient cohorts are not aligned due to the different construction and data completeness/submission requirements for quality measures, for example, MIPS Q#355 includes outpatient surgeries whereas MSPB Clinician is focused on inpatient care. We also examined the potential for analyzing a relationship with MIPS Q#458 All-cause Hospital Readmission, a claims-based outcomes measure evaluating unplanned readmissions. However, there was no publicly available reporting data for this measure in 2017.[7] The measure construction also means

[6] CMS, 2017 Quality Payment Program Experience Report – Appendix https://qpp-cm-prod-content.s3.amazonaws.com/uploads/492/2017%20QPP%20Experience%20Report%20Appendix.zip

[7] CMS, 2017 Quality Payment Program Experience Report – Appendix

that even with complete data, the patient cohort and attributed clinicians would have significant disconnect; for example, MIPS #458 is only reported for groups of 16 or more clinicians meeting a 200 patient case minimum and attributes based on CPT/HCPCS codes denoting primary care services. As such, we were unable to test the empirical validity of a conceptual relationship of the MSPB Clinician measure with MIPS quality measures.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Face Validity

The results of the assessment of face validity indicate that a convened group of experts had high levels of agreement with the measure's ability to provide an accurate reflection of costs, and to distinguish good and poor performance. The survey questions and mean rating for each question are provided below:

Question 1: Indicate the extent to which you agree that these refinements help the measure provide an accurate reflection of the costs related to inpatient care: (i) Attribute medical and surgical episodes using different rules, and (ii) Remove certain services considered out of the reasonable influence of the clinician's care decision from the measure

Response: 12 members agreed (rating between 4-6), 3 members disagreed (rating between 1-3)

Mean Rating[8]: 4.9 out of 6 (somewhat to moderately agree)

Question 2: Indicate the extent to which you agree with the following statement comparing the revised MSPB Clinician measure (in use from MIPS 2020 onwards) to the previous version of the measure (used in MIPS from 2017 to 2019): "The scores obtained from the revised MPSB Clinician measure provide a more accurate reflection of the costs for inpatient episodes of care than the previous version of the measure, and can better distinguish good and poor performance."

Response: 12 members agreed (rating between 4-6), 3 members disagreed (rating between 1-3)

Mean Rating: 4.9 out of 6 (somewhat to moderately agree)

Question 3: Indicate the extent to which you agree with the following statement about the MSPB Clinician measure: "The scores obtained from the revised MSPB Clinician measure as specified will provide an accurate reflection of the costs for inpatient episodes of care, and can be used to distinguish good and poor performance on cost effectiveness."

Response: 14 members agreed (rating between 4-6), 1 member disagreed (rating between 1-3)

Mean Rating: 5.3 out of 6 (moderate to strongly agree)

Question 4: If you disagree with the statement, what aspects of the measure do you believe should be changed for you to agree with the statement?

Response: The member's comment expressed concern about overlapping costs being captured by episode-based cost measures. Other TEP members who agreed with the statement provided general thoughts about the measure, including incorporating Part D costs, further field testing, and accounting for different DRG coding behavior across hospitals.

Empirical Validity

Table 4 present results for the first analysis of validity where the mean O/E cost ratio for all episodes is 1.00. The mean O/E cost ratio for episodes with downstream acute readmission is 1.58, compared with 0.91 for episodes without downstream acute readmission. The mean O/E cost ratio for episodes with PAC is 1.20, while for episodes without PAC is 0.80.

Table 4. Distribution of Observed to Expected Ratios

[8] The mean rating is a simple average. It is calculated by multiplying the number of responses for each rating by the rating, and dividing by the total number of responses.

Cost Driver Category	Observed to Expected Ratios						
	Mean	Std. Dev.	Percentiles				
			10th	25th	50th	75th	90th
All Final Episodes	**1.00**	**0.52**	**0.55**	**0.66**	**0.84**	**1.18**	**1.67**
Episodes with downstream acute (re)admission	1.58	0.66	0.94	1.13	1.41	1.85	2.42
Episodes without downstream acute (re)admission	0.91	0.42	0.53	0.64	0.79	1.02	1.45
Episodes with Post-Acute Care (IRF, LTCH, HH, SN)	1.20	0.56	0.64	0.81	1.06	1.46	1.92
Episodes without Post-Acute Care (IRF, LTCH, HH, SN)	0.80	0.38	0.51	0.60	0.71	0.87	1.14

Table 5 includes results from the clinical themes analysis. There is a weak correlation between the index admission itself (correlation: 0.08) and the risk-adjusted cost. The clinical themes analysis also demonstrated that there is a strong correlation between cost for readmissions (correlation: 0.47) and risk-adjusted cost. Correlation between Outpatient E&M services, procedures, and therapy (correlation: 0.26) and risk-adjusted cost is moderate. Finally, correlation between PAC: SNF (correlation: 0.34) and risk-adjusted cost is moderate to high, while the correlation between another PAC setting, home health (correlation: -0.18), is negative. The correlation between PAC IRF/LTCH (correlation: 0.15) and risk-adjusted cost is moderate to low.[9]

Table 5. Pearson Correlation Statistics between Costs for Clinical Themes with Risk-Adjusted and Expected Costs

Clinical Theme	Average Cost of Grouped Clinical Theme	Pearson Correlation With Risk Adjusted Cost	Pearson Correlation With Predicted Cost
Acute Inpatient Services: Index Admission*	$11,561	0.08	0.87
Acute Inpatient Services: Readmission	$8,863	0.47	0.04
Emergency Services Not Included in Hospital Admission	$739	0.08	-0.01
Outpatient E&M Services, Procedures, and Therapy	$850	0.26	0.01
Post-Acute Care: Home Health	$1,933	-0.18	0.01
Post-Acute Care: IRF/LTCH	$22,518	0.15	0.55
Post-Acute Care: SNF	$11,181	0.34	0.06

*The MS-DRG of the index admission is included in risk adjustment

[9] Conventional standards consider a Pearson correlation of 0.37 or larger to be a large association (Cohen, 1988 and 1992). 1. Cohen J, Statistical Power Analysis for the Behavioral Sciences, 2nd ed., 1988. 2. Cohen J, A power primer, Psychol Bull, 1992;112(1):155-159.

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Face Validity

This measure was assessed by a group of experts. Out of 15 respondents to the survey, 14 (93%) agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost effectiveness.

Empirical Validity

For the first test, as expected, the average O/E cost ratio for episodes with downstream acute readmissions is higher than for episodes without downstream acute readmissions. This result demonstrates that the MSPB Clinician measure is able to accurately capture higher resource use related to readmissions. Similarly, episodes with PAC services (i.e. HH, SNF, IRF, or LTCH) also have a substantially higher average ratio of observed to expected cost than episodes without PAC services.

The second test, the clinical themes analysis, demonstrates that higher risk-adjusted cost is strongly associated with themes related to readmissions and some types of post-acute care services (SNF), and linked – though more weakly – to themes relating to services performed during the episode window and other types of post-acute care services (IRF/LTCH). These results indicate that the measure is able to capture higher cost services, which a cost measure should do to be able to distinguish provider performance.

Below are some more detailed interpretations of the correlations between clinical themes and risk-adjusted cost.

- Since the risk adjustment model adjusts for the MS-DRGs of the index admissions, the correlation between the acute inpatient services performed including and during the index admission and the risk-adjusted cost is quite small.
- The correlation between Outpatient E&M services, procedures, and therapy shows that a larger number of services provided during the episode window is associated with higher observed and expected costs for this clinical theme.
- As expected, episodes that have readmissions have strong correlation with risk-adjusted cost. Hospital readmissions are associated with higher costs and this increase in resource use is captured in the measure by a larger observed over expected cost ratio.
- Between two types of post-acute care services, SNF and home health, SNF is associated with higher cost than home health services. Episodes with SNF after hospitalization would be expected to have higher episode costs even after risk adjustment, since SNF is a more costly form of PAC. Similarly, the correlation results show that a beneficiary receiving home health services is inversely related to the risk-adjusted cost, as a lower intensity form of PAC.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b3*

2b2.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exclusions are used in the MSPB Clinician measure to ensure that the patient population is comparable and as part of data processing. To ensure an appropriate patient cohort within the scope of the measure focus on inpatient hospitalizations/admissions, exclusions focus on removing beneficiaries where fair comparisons cannot be made across providers. Exclusions for data processing are designed to ensure that sufficient data are available to accurately determine episode spending and calculate risk adjustment for each episode. The exclusions for patient cohort comparability, along with their rationales, are listed below:

- Episodes where beneficiary death date occurred before the episode end.
    - Episodes where the beneficiary died may be unusually high-cost, due to perimortem treatment costs, or unusually low-cost, due to the truncated episode window. Neither of these cases accurately reflects the efficiency of the clinician performing the treatment.
- Episodes where the index admission was not performed in an inpatient facility located in the U.S.
    - Episodes with admissions in hospitals located outside the 50 states and D.C. are excluded to remove episodes in which the beneficiary may receive care not reimbursed under Medicare Parts A and B.
- Episodes in which the index admission occurred in a non-Acute Hospital or in a Critical Access Care (CAH) hospital
    - Episodes where the beneficiary's hospitalization did not occur in an acute care hospital facility are excluded from the measure calculation. These episodes are excluded from the measure to ensure that only costs incurred in facilities paid under the IPPS are considered. Medicare pays for the same inpatient and outpatient services provided at CAHs as for those provided at IPPS acute care hospitals. However, while hospitalizations at IPPS hospitals are paid based on prospectively set rates (where payment is based on the average resources used to treat Medicare patients with different Diagnosis-Related Groups), CAH payments are based on each CAH's costs. Therefore, CAHs have an ability to significantly influence Medicare payments they receive and, therefore, cannot be compared to IPPS hospitals in a fair manner.
- Episodes where the index admission either begins as a transfer from another hospital or ends as a transfer to another hospital
    - Depending on the timing and nature of the transfer, it is difficult to differentiate the level of involvement between the clinician(s) providing care in the initial facility and the clinician(s) of the receiving facility. Since we cannot differentiate each clinician's role and amount of involvement with the patient, these episodes are excluded to ensure comparability with other episodes.

Given the rationales for these exclusions, we would expect these excluded episodes to have a different risk profile than the included episodes, such as a higher or lower mean cost, or a different distribution of costs (e.g., a long tail of high-cost episodes). For the exclusions, we examined the number of episodes and beneficiaries affected, as well as the distributions of observed cost and ratio of observed over expected spending (calculated by applying existing risk factor coefficients to the excluded episodes) for excluded episodes. We then compared the cost characteristics of the excluded episodes to those of episodes remaining after the described exclusions to assess the distinctness between the two patient cohorts.

Please also see Section 2b6 (*Missing Data Analysis and Minimizing Bias*) of this testing form for more information on exclusions implemented as part of data processing and completeness requirements.

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Table 6 below presents the percentage of episodes captured by each exclusion, observed cost statistics, and observed over expected (O/E) cost ratios for the MSPB Clinician measure exclusions. Cost statistics are also provided for the remaining set of episodes after the described exclusions are applied for comparison. Appendix Table 2b2.2 provides more detailed cost distributions for measure exclusions.

Table 6. Cost Statistics for Measure Exclusions

Exclusion	Episodes		Observed Cost			O/E		
			Mean	Percentile		Mean	Percentile	
	#	%		10th	90th		10th	90th
All Episodes Meeting Triggering Logic	**10,658,462**	**100.0%**	**$21,702**	**$6,612**	**$41,075**	**1.03**	**0.51**	**1.77**
Episodes in which Inpatient Stay had Transfers or Death Discharge Status Codes or episodes that overlapped with an IP Stay with Transfer or Death Discharge Status Codes	753,178	7.1%	$32,053	$9,422	$63,635	1.23	0.45	2.44
Episodes in which beneficiary Death occurred within 30 Days Post Discharge	982,827	9.2%	$23,956	$8,298	$45,346	0.87	0.41	1.57
Episodes in which Inpatient stay occurred in a non-Acute Hospital or in a Critical Access Care (CAH) hospital	1,212,822	11.4%	$26,779	$6,605	$51,078	1.39	0.50	2.47
Episodes with Inpatient Facility located in Excluded Regions	29,817	0.3%	$14,787	$4,628	$30,728	0.79	0.43	1.27
Remaining Episodes	**6,294,955**	**59.1%**	**$20,112**	**$6,606**	**$37,540**	**1.00**	**0.53**	**1.66**

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The statistical results indicate that the excluded episodes have different mean observed episode costs, which may be due to variation in payment that is not under the influence of the attributed clinician. These episodes can also be excluded due to clinical considerations to ensure a comparable patient cohort that will yield meaningful information to attributed clinicians. Further discussion of the results for each exclusion is provided below.

Episodes ending in death or transfer: The results indicate a large difference between the mean observed episode costs for episode in which the inpatient stay ends in a transfer or death at $32,053 compared to episodes not excluded at $20,112. Similarly, episodes in which a death occurs in the post-discharge period have a higher observed cost at $23,956. Furthermore, the ratio of observed to expected episode cost for episodes with index admissions with discharges related to transfers or death is 1.23 indicating that the episodes are more costly than is predicted by the risk adjustment model's consideration for patient characteristics and condition observed prior to the admission. For episodes in which the beneficiary dies within 30 days after discharge, the mean O/E cost ratio predicts the episodes as being more expensive based on the patient's characteristics prior to the episodes start than observed with a mean value of 0.87. This could be due to the episode window being truncated by the death of the patient, resulting in less cost being assigned. Including episodes with inpatient stays ending in death or transfer in the measure calculation may distort measure scores where patients require more resource use than can be predicted based on the patient case-mix upon admission and would give the appearance of less cost effective care. Inversely, when death occurs after discharge, including these episodes may skew a provider performance to look more efficient.

Episodes in which Inpatient stay occurred in a non-Acute Hospital or in a Critical Access Care (CAH) hospital: Since CAHs are paid under a different payment system and have a greater impact on the amount of Medicare reimbursement they receive, they have a much higher O/E ratio than acute care hospitals paid under IPPS (1.39 versus 1.03, respectively). Similarly, the difference in the mean observed cost of these episodes and episodes not excluded is $6,667 more. Since the risk adjustment model does not account for this difference in cost related to the facility type, these cases are not included in the measure to ensure that only costs reimbursed under similar payment systems are considered.

Episodes with Inpatient Facility located in excluded regions: Episodes with inpatient facilities in the following regions are excluded from the measure calculation because hospitals in those areas are not paid under the inpatient prospective payment system (IPPS): Puerto Rico, Virgin Islands, Canada, Mexico, Samoa, Guam, Marinas, and foreign countries.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

2b3.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with** 109 **risk factors**

☒ **Stratification by** 26 **risk categories**

☐ **Other,** Click here to enter description

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

Differences in case mix are controlled for using a statistical risk model with 109 risk factors. The risk adjustment model for the MSPB Clinician measure broadly follows the CMS-HCC risk adjustment methodology used in the Medicare Advantage (MA) program. Severity of illness is measured using HCCs, indicators of enrollment and long-term care status, and disease interactions. There also 12 categorical age variables included in the model.

The model includes 79 HCC indicators derived from the beneficiary's Parts A and B claims during the period 90 days prior to the episode start date and are specified in the CMS-HCC Version 22 (V22) 2016 model. Episodes for beneficiaries without a full 90-day lookback period are excluded from the measure. This 90-day period is used to measure beneficiary health status and ensures that each beneficiary's claims record contains sufficient data for risk adjustment purposes.

In addition, the risk adjustment model includes status indicator variables for whether the beneficiary qualifies for Medicare through Disability or has ESRD. The model also includes an indicator of whether the beneficiary recently required long-term care, defined as 90 days in a long-term care facility without being discharged to community for 14 days. Beneficiaries who need to reside in long-term care facilities typically require more intensive care than beneficiaries who live in the community. These enrollment and long-term care status variables are non-diagnostic based indicators of severity of illness.

The model also accounts for disease interactions between HCCs and/or enrollment status variables included in the MA model. These interactions are included because certain combinations of comorbidities increase costs more than is predicted by the HCC indicators alone.

As with the CMS-HCC model, the risk adjustment approach for this measure uses an ordinary least squares linear regression model. The predicted, or expected, cost is winsorized at 0.5th percentile to make sure episodes with unusually small predicted cost, which would lead to abnormally large O/E ratios, do not dominate certain clinicians' final score. The winsorized expected costs are renormalized to ensure the average expected episode cost is the same before and after winsorizing. Then, extremely low- or high-cost outlier episodes with residuals below the 1st percentile or above the 99th percentile are excluded to reduce the effect of these episodes that deviate the most from their expected values in absolute terms. The expected cost after excluding these outliers is again renormalized to ensure that average expected costs are the same after outlier removal.

Finally, the risk adjustment model outlined above is performed separately for set of episodes within each MDC as determined by the MS-DRG of the index admission.

Full details of the risk adjustment model are in the Measure Codes List File (linked in Section S.1). Appendix Table 2b3.6.b provides regression coefficients, standard errors and other statistics for each model.

2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

Clinical Factors: The CMS-HCC model was selected 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 and is calibrated on Medicare fee-for-service beneficiaries. 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) and is exhaustive on these code sets. Because the CMS-HCC model has already been extensively tested and is used for a large Medicare Part C population, we focus our testing on how the CMS-HCC model was adapted to the MSPB Clinician measure.[10,11,12]

The statistical risk model is estimated separately for each MDC, which is determined by the MS-DRG of the index admission; in turn, these are generally grouped according to principal diagnoses or major procedures. This

[10] In 2018, 20 million beneficiaries were enrolled in Medicare Part C plans and incurred $230 billion to cover Medicare Part A and Part B services for Medicare Advantage enrollees (MEDPAC Data Book *Healthcare Spending and the Medicare Program*, June 2019, http://www.medpac.gov/docs/default-source/data-book/jun19_databook_entirereport_sec.pdf?sfvrsn=0)

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

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

risk stratification by MDC is to ensure that the wide range of inpatient care and the different clinical factors that affect resource use are accounted for in the model. Each MDC corresponds to an organ system (e.g., MDC 2 covers diseases and disorders of the eye) or cause for admission (e.g., MDC 22 comprises MS-DRGs related to burns).

The measure also includes a Prior Inpatient Admission risk adjustor to ensure fair comparison between episodes with and without prior inpatient admissions. Episodes where an inpatient stay occurs in 30 days prior to the episode trigger are considered re-admissions that tend to be riskier and more resource-intensive than admissions. However, a risk adjustor for prior inpatient admission ensures that riskier health status of the beneficiary and higher resource use associated with those episodes are being adjusted for and allows us to fairly compare these episodes to other episodes in the population.

We also sought extensive input from clinical experts with experience providing care in inpatient settings throughout the measure re-evaluation process, as described in Section 2b1.2. We also received feedback from the broader stakeholder community through field testing and the pre-rulemaking and rulemaking processes. Convened panels of experts provided input on other clinical factors for the measure to account for: (i) attributing episodes differently for surgical and medical MS-DRGs, and (ii) excluding services for each MDC that are clinically unrelated to inpatient care. Excluding clinically unrelated services from the episode cost reduces variation unlikely to be under the influence of the clinicians care, improving the accuracy of the model.

Social Risk Factors: According to a 2014 National Quality Forum report[13], the mechanisms underlying differences in resource use by socioeconomic status and race are complex and may be impacted by factors such as financial resources, community resources, historical and current discrimination, and reduced access to preventive services. Provider assumptions or implicit biases may impact quality of care for beneficiaries of different races. These factors may result in inefficient care, increased disease severity, or greater morbidity, leading to higher Medicare spending for beneficiaries depending on socioeconomic status or demographic status. Other social risk factors identified by the literature that can affect resource use include income, insurance (e.g., Medicaid), education, race and ethnicity, sex, social relationships, and residential and community context including rurality.[14,15,16]

Given the conceptual relationship between these social risk factors and resource use, we analyzed the impact of the following beneficiary-level and Census-Block Group-level social risk factors: income, education, employment, race, sex, dual status, and AHRQ Index. These factors are also listed in Section 1.8.

We used the CMS Enrollment Database (EDB), and Common Medicare Environment (CME) to determine dual eligibility, race, and sex. Socioeconomic status was determined by two approaches: a) using income, education and employment status as categorical dependents and b) using Agency of Healthcare Research and Quality (AHRQ) SES Index as a continuous dependent. Both approaches used data from the 2017 American Community Survey (5-year file) by linking episodes to census block groups, and ZIP code when census block group is missing.

Social risk factors were examined relative to the base model set of risk adjustment variables from the CMS-HCC V22 2016 model, disability status, ESRD status, interaction variables, and recent long-term care use, and in a step-wise fashion to determine the potential value of each social risk factor considered.

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

---

[13] National Quality Forum. "Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors: Technical Report." National Quality Forum: August 2014

[14] National Academies of Sciences Engineering and Medicine (U.S.). Committee on Accounting for Socioeconomic Status in Medicare Payment Programs, Kwan LY, Stratton K, Steinwachs DM. Accounting for social risk factors in medicare payment : a report of the National Academies of Sciences, Engineering, Medicine. Washington, DC: The National Academies Press; 2017

[15] 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

[16] Medicare Payment Advisory Commission. Beneficiaries Dually Eligible for Medicare and Medicaid. 2018

☒ **Published literature**

☒ **Internal data analysis**

☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

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., ACOs, previous physician QRUR programs, and other measures such as NQF #2158: 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 Pope et al (2011) report and the December 2018 CMS Report to Congress on risk adjustment in Medicare Advantage.[17,18]

Appendix Table 2b3.6.b includes regression coefficients and standard errors for each of the covariates used in the risk adjustment model on the measure's specific population.

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

We analyzed race, sex, dual status, income, education, and unemployment as social risk factors (more information on these variables can be found in Section 1.8 of this document). Beneficiary sex and dual status were obtained from the EDB and CME. Information on income, education, and unemployment was obtained from ACS data. Beneficiaries without geographic information necessary to obtain ACS data where excluded, approximately 1.6 percent of beneficiaries[19].

The percentage of female beneficiaries range from 27.5 percent to 63.8 percent across the 23 of the 26 MDCs in this measure that reasonably occur for both sexes (MDC 13 and MDC 14 are nearly 100 percent female as they are related to pregnancy, childbirth, and the female reproductive system, while MDC 12 is 0 percent female as it is related to the male reproductive system). For 22 out of 26 MDCs, the majority of the beneficiaries (57.8% - 84.4%) have non-dual status. The MDCs with a minority of non-dual status beneficiaries includes MDC 14 – Pregnancy, Childbirth, and the Puerperium (8.6%), MDC 25 – Human Immunodeficiency Virus Infections (28.5%), MDC 19 – Mental Diseases and Disorders (43.7%), and MDC 20 – Alcohol/Drug Use and Alcohol/Drug Induced Organic Mental Disorders (49.7%). Income level is categorized into high, medium, and low from the continuous average income variable in ACS; therefore, each category has 33.3 percent of episodes. While 1.9 to 7.7 percent of beneficiaries across all MDCs are classified as having below a high school education level, between 92.3 and 98.1 percent of beneficiaries across MDCs are classified at a high school level or greater. Finally, 16.6 to 37.5 percent of beneficiaries have high unemployment designation (>10% for the Census Block Group).

We examined the impact of including social risk factors into our risk adjustment model by running goodness of fit tests when different risk factors are added and compared to the base risk adjustment model, where the base risk adjustment model refers to the full standard set of risk adjustment variables from the CMS-HCC V22 2016 model, disability status, ESRD status, interaction variables, and recent long-term care use. We ran a step-wise

---

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

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

[19] Due to this exclusion, coefficients and model fit presented for the base model analyzed within the SRF testing will slightly differ to those presented for the model testing conducted in Section 2b3.5.

regression to include the following additional social risk factors on top of the adapted base CMS-HCC model (Model 1):

- Model 2: sex
- Model 3: dual status
- Model 4: sex + dual status
- Model 5: sex + dual status + race
- Model 6: sex + dual status + income + education + unemployment
- Model 7: sex + dual status + AHRQ SES Index
- Model 8: sex + dual status + race + income + education + unemployment
- Model 9: sex + dual status + race + AHRQ SES Index

The step-wise regressions help evaluate individual as well as joint significance of the social risk factors. We examined the impact of including social risk factors into our risk adjustment model with T-test of individual significance and F-test of joint significance.

First, we analyzed the model coefficients and p-values for each of the base and social risk factor models to understand whether any of the social risk factor covariates are predictive of episode cost. The T-test and F-test revealed significant p-values varied across the stratification of MDC, indicating that social risk factors are likely predictive factors for determining resource use among beneficiaries for the relevant characteristic and MDC. For example, the AHRQ SES Index has a p-value less than or equal to 0.05 for 10 of the 26 MDC stratifications. The analysis also shows that the directions of the effects of social risk factors are not consistent. For example, low income episodes (as compared to high income episodes) and the AHRQ SES index may display both significant positive and negative coefficients of spending across MDCs. Considering the categorical factor low income as an example, positive coefficients for low income may indicate that people with lower income tend to be more vulnerable and are in need of additional resource use in their care. On the contrary, negative coefficients could indicate lower income people are expected to spend less, which may be a result of low income patients having financial incentives to use less health care resources. They may be burden by co-pays for other services that they received covered by Medicaid. Due to the inconsistency across coefficients, it is unclear what underlying confounding traits are be adjusted and how incentives for clinicians to be selective of patient and/or the amount of resource provided to patients based on socioeconometric and sociodemographic status could be perversely affected. Appendix Tables 2b3.4b.a and 2b3.4b.b present these results.

Secondly, we analyzed the impact of adding social risk variables on overall model performance by looking at the differences in the O/E cost ratio with and without social factors in the risk adjustment model. When including social risk factors in our risk adjustment regression, minor differences in the O/E ratios, even for providers at high or low extremes of risk, indicates that social risk factor effects on the model performance are likely captured through existing risk adjustment variables. When including sex, dual, race and SES as categorical variables (i.e. income, education, and unemployment), the ratio of observed over expected costs for 90.6 percent of TINs and 92.3 percent of TIN-NPIs changed by less than ±0.01. At a higher threshold, 99.7 percent of TINs and 99.8 percent of TIN-NPIs changed by ±0.03 or less. Using the AHRQ SES index in place of the categorical variables for income, education, and unemployment yielded similar results. Appendix Table 2b3.4b.c presents these results in detail.

Finally, we analyzed the correlation between measure scores calculated with and without the social risk factors. The measure scores calculated with and without these social factors were highly correlated at both the TIN level (Spearman correlation coefficient of 0.998), and the TIN-NPI level (Spearman correlation coefficient of 0.999). These results indicate that the inclusion of social risk factors in the current risk adjustment model would have a limited effect on measure scores. Appendix Table 2b3.4b.d presents these results in detail.

Due to the inconsistent direction and limited impact of social risk factor effects under the current risk adjustment model, we believe the MSPB Clinician measure risk adjustment model sufficiently accounts for the effects of social risk factors on clinician measure scores.

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*). *Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.
If stratified, skip to 2b3.9

To analyze the validity of the current risk adjustment model, we examined three analyses: (1) R-squared and adjusted R-squared for the regression models, (2) predictive ratios and O/E cost ratios to examine the fit of the models at different levels of patient complexity, and (3) coefficient estimates, standard errors, and p-values for each MDC.

1) *R-squared and adjusted R-squared* were calculated for each MDC. The results should be evaluated in the context of the service exclusion rules for each MDC, which indicate which costs are counted in the measures and which costs are not counted. This is a distinction from true all-cost measures, as a low R-squared does not necessarily indicate that a measure reflects variation unrelated to clinical care, while a high R-squared does not necessarily indicate the opposite; instead, the risk adjustment models must be evaluated in concert with the service exclusion rules.
2) *Predictive ratios and O/E cost ratios* were calculated for each "risk decile" for the MDC. A "risk decile" is based on the risk scores, which indicate how costly episodes are expected to be, as predicted through risk adjustment. After arranging episodes into deciles based on their risk score, we calculated the predictive ratios and average O/E cost ratios for each decile. The predictive ratio aims to examine the fit of the model at different levels of patient complexity to examine the model's ability to predict both very low and high cost episodes, and is calculated using the formula of average (expected cost)/average (observed cost) for all episodes in each decile. Similarly, the O/E cost ratio demonstrates the model's prediction accuracy, and is calculated using the formula of average (observed cost/expected cost) for all episodes in each decile.
3) *Coefficient estimates, standard errors, and p-values* were run for each MDC to consider the extent to which the coefficients for the risk factors are predictive of episode cost. Results for individual risk adjustment variables should be viewed in the context of the entire model and set of MDCs, rather than being analyzed individually. For instance, coefficients indicate the incremental effect of a model variable, holding all other variables fixed. As another example, interactions between model variables must be interpreted in concert with the effects of those variables in isolation.

Results and interpretation of these analyses are discussed below in Sections 2b3.6-2b3.10.

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

The range of R-squared values for the MSPB clinician cost measure risk adjust models, calculated by dividing explained sum of squares by total sum of squares ranges between 0.09 – 0.64 across the MDCs. The adjusted R-squared range is 0.09 – 0.63.

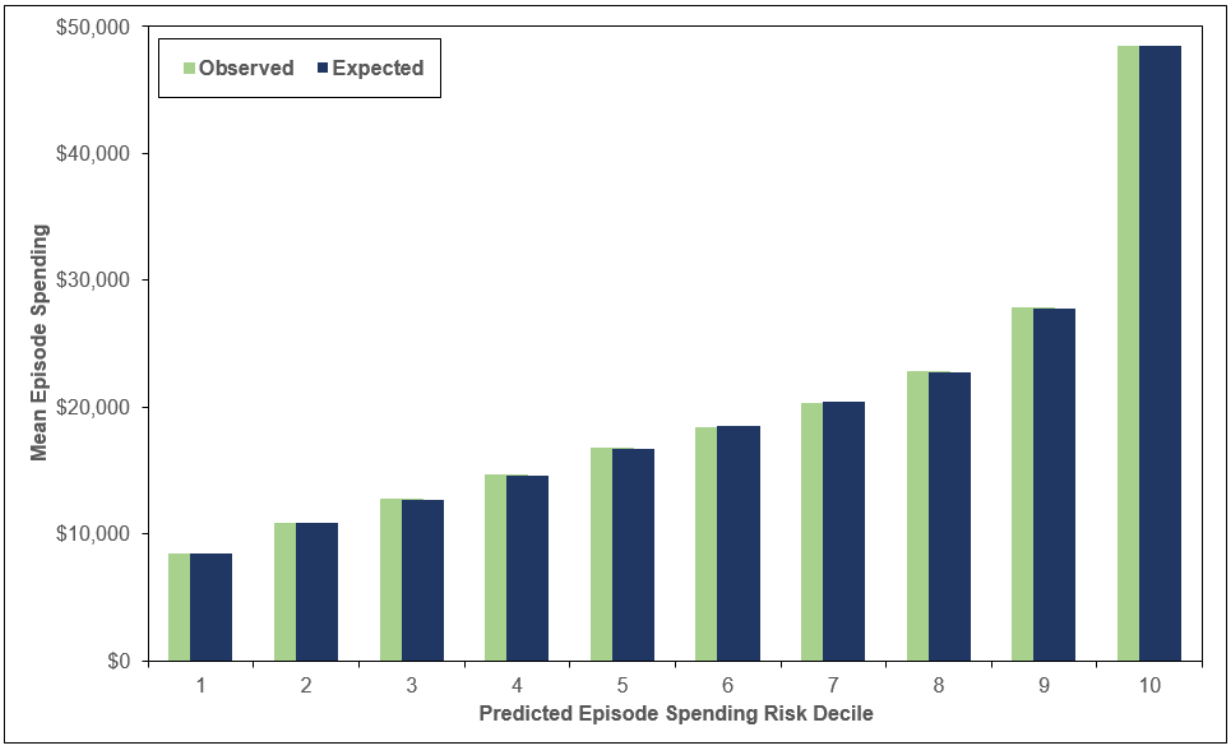
Appendix Table 2b3.6.a provides the R-squared and adjusted R-squared values for each risk adjustment model. Appendix Table 2b3.6.b provides regression coefficients, standard errors and other statistics for each model. More information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.[17]

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

We interpret calibration as how accurately the risk model's predictions match the actual episode cost. We calculate the average O/E cost ratio for each risk decile to demonstrate the model's prediction accuracy for both high and low cost episodes. **Figure 1** presents the comparison for each decile and shows the observed to expected costs to be close, with a difference less than 2 percentage points. The average observed to expected

cost is generally close to one, 0.99 to 1.01, across risk deciles, indicating that the model is accurately predicting actual episode cost across risk deciles. Full results can be seen in Appendix Table 2b3.7.a

Figure 1. Risk Adjustment Model Diagnostics: Comparison of Observed and Expected Spending by Predicted Spending Risk Deciles



2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Analysis of predictive ratios by risk decile for the measure shows that the model has consistent predictive ratios across risk score deciles, with each decile having a predictive ratio between 0.99 and 1.01. Full results can be seen in Appendix Table 2b3.7.a.

2b3.9. Results of Risk Stratification Analysis:

N/A

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
As demonstrated in Section 2b3.7 and 2b3.8, the average O/E cost ratios and the predictive ratios for all risk deciles are close to one. These results indicate that the model is accurately predicting spending, regardless of overall risk level. There was no evidence of excessive under- or over-estimation at the extremes of episode risk.

The R-squared values for the model, which measure the percentage of variation in results predicted by the model, are above 0.20 for 21 of the 26 MDCs with an average value of 0.3. As noted in Section 2b3.5, these results should be interpreted alongside service exclusion rules, which remove clinically unrelated services, so the resulting variation is reflective of variation related to factors within a clinician's sphere of influence. The service exclusions improve the actionability of the measure while potentially reducing its fit statistics (adjusted R-squared). Unrelated services were purposefully and carefully excluded to improve the ability to interpret and compare MSPB Clinician scores across providers. Since excluded services outside the influence of the attributed clinician may be well predicted by patient risk factors, excluding them can reduce the explained portion of the cost variance and the model's adjusted R-squared.

2b3.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

We used two methods to identify statistically significant and meaningful differences in the MSPB Clinician measure scores. The purpose of these analyses is to ensure that there is a sufficiently large difference in measure scores among clinicians to discern a meaningful difference in performance. First, we analyze the distribution of measure scores for clinicians defined by these meaningful characteristics, as well as for the overall measure. We stratified the measure scores by provider characteristics to confirm that the measure behaves as expected for different types of clinicians. Stratification is performed for each of the following characteristics: urban/rural, census division, census region, risk score, and the number of episodes attributed to the clinician. In our second test, 95% confidence intervals (CI) were calculated using the variance of the provider mean. We then compared each clinician's 95% CI to the national average measure score to determine if the clinician's performance was significantly different from the national mean.

2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

MSPB Clinician scores are distributed fairly symmetrically and have a good deal of variability. For TINs, the standard deviation is $1,785, and 99/1, 90/10, and 75/25 percentile ratios are 1.60, 1.25, and 1.11, respectively. For TIN-NPIs, the standard deviation is $1,885, and 99/1, 90/10, and 75/25 percentile ratios are 1.58, 1.26, and 1.13, respectively. **Figures 2 and 3** display the distribution of the MSPB ratios (i.e. TIN/TIN-NPI's mean observed to expected cost across attributed episodes), a direct scalar of their measure score.

Figure 2. Distribution of MSPB Clinician Ratio for TINs

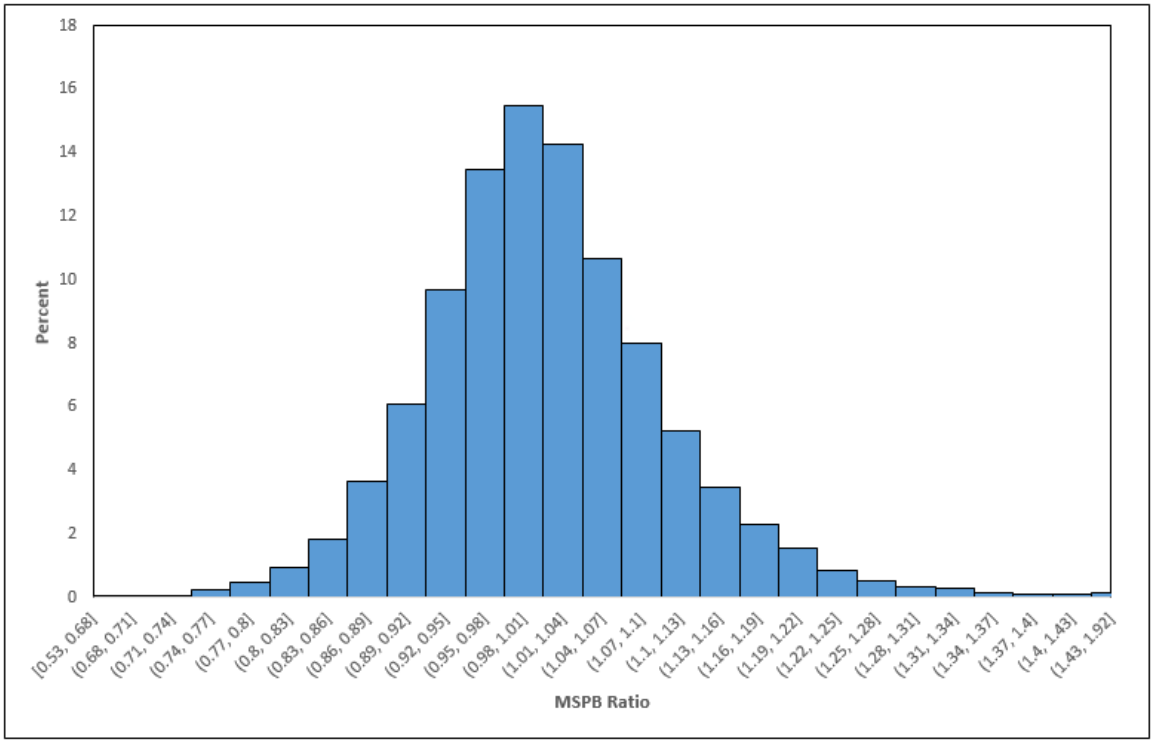
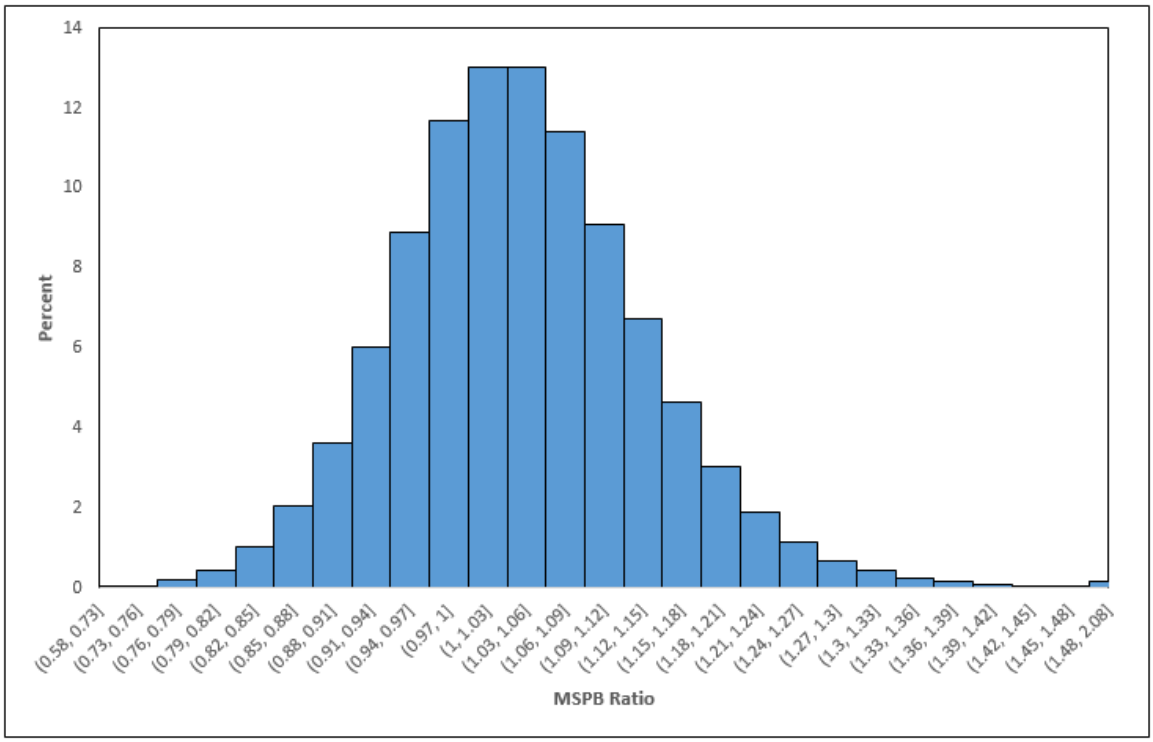


Figure 3. Distribution of MSPB Clinician Ratio for TIN-NPIs



The results indicate the measure is capturing differences in performance measure scores. The results also show that there are not systemic differences in clinician performance by provider characteristic. For instance, the difference in the mean scores for clinicians across nine census divisions (excluding 'Unknown') are within less than $950 for both the TIN and TIN-NPI testing (i.e., $18,420 - $19,639 at the TIN level and $19,298 - $20,227 at the TIN-NPI level). Similarly, clinicians in urban areas seem to perform comparably to those in rural areas with less than a $580 difference in mean score for both reporting levels.

Analysis of clinicians by number of episodes indicates that clinicians with more episodes perform similarly to those who have fewer episodes. We also analyzed clinicians by risk score decile, as variation by risk score decile could indicate that the risk adjustment model is over- or under-correcting for clinicians with systematically riskier patients. Measure scores also show little variation by risk score decile, with a range in mean TIN score of $18,129 to $19,751 and a range in mean TIN-NPI score of $18,454 to $20,366, indicating that the risk adjustment model is overall functioning as intended. Full results are provided in Appendix Table 2b4.2.

Due to the high level of reliability of the MSPB Clinician scores, demonstrated in Section 2a2, small differences in scores can be interpreted as meaningful. This is confirmed by our analysis of statistical significance: 12.4 percent of TINs and 8.0 percent of TIN-NPIs had scores that were statistically significantly higher than the national mean, while 20.0 percent of TINs and 15.3 percent of TIN-NPIs had scores that were statistically significantly lower (**Table 7**).

Table 7. Proportion of Measure Scores Statistically Significantly Different From the National Average

Provider Level	# of Providers	Statistically significantly lower than national mean		Not statistically significantly different from national mean		Statistically significantly higher than national mean	
		#	%	#	%	#	%
TIN	19,213	3,835	20.0%	12,996	67.6%	2,382	12.4%
TIN-NPI	126,628	19,326	15.3%	97,138	76.7%	10,164	8.0%

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

There is clinically and practically significant variation in MSPB Clinician measure scores, indicating the measure's ability to capture differences in performance. Our findings regarding variation in measure scores are consistent with expert clinician input and a high face validity rating from expert clinicians that that scores obtained from the measure specification will provide an accurate reflection of the costs for inpatient episodes of care, and can be used to distinguish good and poor performance on cost effectiveness (see Section 2b1.2). For example, empirical results indicate that the measure is appropriately accounting for different risk profiles of patient case-mix. Further, the measure is performing comparably across clinicians with different characteristics, such as geographic location and rurality. These suggest that differences in scores are due to meaningful differences in performance, rather than patient or clinician effects. In this way, the measure can capture meaningful differences in resource use and, thus, provide actionable feedback to clinicians on how to improve their performance through care practice changes.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of*

specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A

2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
N/A

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Since the MSPB Clinician measure is calculated using Medicare claims data, we expect a high degree of data completeness. To ensure further that we have complete and accurate data for each beneficiary who opens an episode, we excludes episodes where beneficiary date of birth information (an input to the risk adjustment model) cannot be found in the EDB or the beneficiary death date occurs before the episode trigger date.

The MSPB Clinician measure also excludes episodes where the beneficiary is enrolled in Medicare Part C or has a primary payer other than Medicare in the 90-day lookback period and episode window. In such situations, Medicare Parts A and B claims data may not capture the complete clinical profile for the beneficiary needed to capture the clinical risk of the beneficiary in risk adjustment. Furthermore, Parts A and B claims data may not capture all Medicare resource use if some portion of the beneficiary's care is covered under Medicare Part C.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
Table 8 below presents the frequency of missing data across the four categories of missing data, which caused episodes to be excluded from the MSPB Clinician measure. Frequency is presented in terms of the number of episodes excluded due to missing data, as well as the number of TINs and TIN-NPIs who had at least one episode excluded due to missing data. The missing data categories are:

- Beneficiary date of birth is missing
- Beneficiary death date occurred before the trigger date
- Beneficiary has a primary payer other than Medicare during the episode window or in the 90-day lookback period
- Beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 90-day lookback period and episode window

- **Table 8. Missing Data Categories for the MSPB Clinician Measure**

Exclusion	# Episodes	# TINs	# TIN-NPIs
Missing birth date	*	*	*
Death before trigger	*	*	*
Primary payer other than Medicare	1,132,724	40,140	308,645
Not continuously enrolled in Parts A and B	1,543,418	40,174	328,166

*denotes that there were fewer than 11 episodes

Additional descriptive statistics for the episodes are provided in Appendix Table 2b6.2 and 2b2.2.

2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

As the MSPB Clinician measure is calculated with Medicare claims data, we expects a high degree of data completeness, which is supported by the limited frequency of missing data for birth date and invalid beneficiary death date information above. Additionally, the measure removes beneficiaries that may have gaps in the Medicare claims history due to alternate enrollment. This data processing step ensures that we have complete and accurate information needed to calculate the measure.

Feasibility

F.1. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

F.1.1. Data Elements Generated as Byproduct of Care Processes.

Generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

F.2. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

F.2.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*)

ALL data elements are in defined fields in a combination of electronic sources

F.2.1a. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.

F.2.2. <u>If this is an eMeasure</u>, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.

Attachment:

F.3. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

F.3.1. Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

Lessons and associated modifications are categorized into three types: data collection procedures, handling of missing data, and sampling data associated with beneficiaries who died during the measurement period.

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 is not practical to wait until all claims for a given month are finalized before calculating this measure. As such, there is 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 three months after the end of the calendar year to collect data for development and testing purposes.

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 was not enrolled in Medicare Parts A and B for the 90 days prior to the episode start date are not included in this measure. This enables the risk adjustment model to adjust accurately for the beneficiary's comorbidities using data from the previous 90 days of Medicare claims. Additionally, the risk adjustment model includes a categorical variable for beneficiary age bracket, so episodes for which the beneficiary's date of birth cannot be located are not included in this measure.

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 does not include episodes for which the beneficiary's date of death occurs prior to the end of the episode window.

F.3.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, and algorithm)?

N/A.

F.3.3. If there are any fees associated with the use of this measure as specified, attach the fee schedule here. (Save file as: F3_3_FeeSchedule)

Usability and Use

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

U.1.1. Current and Planned Use

Specific Plan for Use	Current Use (for current use provide URL)
	Payment Program Quality Payment Program Merit-based Incentive Payment System https://qpp.cms.gov/mips/overview

U.1.2. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Program Name: Quality Payment Program (QPP) Merit-based Incentive Payment System (MIPS)
Sponsor: CMS
Purpose: The Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) established the Quality Payment Program. Under the Quality Payment Program, clinicians are incentivized to provide high-quality and high value care through Advanced Alternate Payment Models (APMs) or the Merit-based Incentive Payment System (MIPS). MIPS eligible clinicians will receive a performance-based payment adjustment to their Medicare payment. This payment adjustment is based on a MIPS final score that assesses evidence-based and practice-specific data across the following categories:

1. Quality
2. Improvement activities
3. Promoting interoperability
4. Cost

As specified in the CY 2020 Physician Fee Schedule final rule (84 FR 62959 through 62979), this measure will be implemented as part of MIPS beginning in the 2020 MIPS performance year and 2022 MIPS payment year.
Geographic Area: U.S.
Number/Percentage of Accountable Entities:
The number of clinicians in the Quality Payment Program varies by performance period. For 2018, there were 889,995 MIPS eligible clinicians receiving a MIPS payment adjustment. [1] As clinicians have choices on how to participate in the Quality Payment Program (e.g., through MIPS or the Advanced APMs, as groups or individuals), the exact number and percentage of clinicians who will receive a performance score on this measure will only be confirmed after the end of each performance period.
[1] CMS, 2018 Quality Payment Program (QPP) Performance Results, https://www.cms.gov/blog/2018-quality-payment-program-qpp-performance-results.

U.1.3. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A.

U.1.4. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3

years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

N/A.

U.2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation. How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Development: TEP and MSPB Service Refinement Workgroup

During measure re-evaluation, Acumen incorporated input from (i) a technical expert panel (TEP) and (ii) the MSPB Service Refinement workgroup. Members of the TEP and the workgroup were selected based on their experience, following separate public calls for nominations. The TEP was composed of 19 members with expertise in cost measure development and evaluation and quality improvement from diverse backgrounds, including clinicians, healthcare providers, academia, and patient advocacy organizations. TEP members provided high-level guidance and initial input on direction of refinements, focusing on attribution and service refinements. The targeted MSPB Service Refinement workgroup was composed of 25 members (including four TEP members) with experience and expertise in a broad range of inpatient care, affiliated with 21 specialty societies. The workgroup provided detailed clinical input on service exclusion rules.

Development: Field Testing

Acumen and CMS conducted a national field testing of 11 episode-based cost measures and two population-level cost measures, including the MSPB Clinician measure, developed during 2018 for a 35-day comment period (October 3, 2018 to November 5, 2018). We provided MSPB Clinician Field Test Reports to a sample of eligible clinician groups and clinicians. Each report included information for the MSPB Clinician measure if the clinician or clinician group was attributed 35 or more episodes. [1] The testing sample was selected to balance coverage and reliability, since a key goal of field testing was to test the measure with as many stakeholders as possible. The number of field test reports shared with the public was:

- Total reports: 793,842
- Total MSPB Clinician reports: 148,382
- TIN reports: 20,852
- TIN-NPI reports: 127,530

All stakeholders, including those who did not receive a field test report, could review a mock field test report that was posted on the CMS website. Other public documentation posted during field testing included: measure specifications (comprising a Draft Cost Measure Methodology document and a Draft Measure Codes List file), a Frequently Asked Questions document, and a Fact Sheet. [2] During field testing, Acumen conducted education and outreach activities, including a national webinar, office hours with specialty societies, and Help Desk support.

Implementation: Pre-Rulemaking and Rulemaking

The MSPB Clinician measure was implemented in MIPS after going through the pre-rulemaking process and notice-and-comment rulemaking. The measure was submitted to and included in the 2018 Measures Under Consideration (MUC) List. It was then considered by National Quality Forum (NQF)'s Measure Applications Partnership (MAP) Clinician Workgroup and Coordinating Committee in December 2018 and January 2019, respectively.

The measure with the revised specifications was proposed for use in the MIPS cost performance category in the CY 2020 Physician Fee Schedule proposed rule. [3] A National Summary Data Report containing information about the measure performance (e.g., measure score distributions by different provider characteristics) was also publicly posted. [4] Stakeholders submitted comments on the proposed rule during a 44-day public

comment period. CMS considered these comments and finalized the measure for use in MIPS from the CY 2020 performance period onwards in the CY 2020 Physician Fee Schedule final rule. [5]

[1] The field test reports were available for download from the CMS Enterprise Portal: https://portal.cms.gov/wps/portal/unauthportal/home/.

[2] The Measure Development Process, Frequently Asked Questions, and Fact Sheet documents are posted on the MACRA Feedback Page: https://www.cms.gov/Medicare/Quality-Payment-Program/Quality-Payment-Program/Give-Feedback.

[3] The CY 2020 Physician Fee Schedule proposed rule can be found here: https://www.federalregister.gov/documents/2019/08/14/2019-16041/medicare-program-cy-2020-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other.

[4] CMS, "National Summary Data Report: 11 Episode-Based Cost Measures and Two Revised Cost Measures, Updated Following Field Testing (Oct-Nov 2018)," MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2018-national-summary-data-report.zip.

[5] The CY 2020 Physician Fee Schedule final rule can be found here: https://www.federalregister.gov/documents/2019/11/15/2019-24086/medicare-program-cy-2020-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other.

U.2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

TEP and MSPB Service Refinement Workgroup

The TEP convened to discuss this measure at two meetings in August 2017 and May 2018, where the panel provided high-level guidance and input on direction of refinements as well as more detailed input on specific approaches to refining attribution methodology and creating service exclusion logic using empirical testing. To determine service exclusions, the TEP recommended convening a targeted workgroup which would become the MSPB Service Refinement workgroup.

The MSPB Service Refinement workgroup attended two webinars in June and July 2018, where members reviewed and discussed empirical analyses and used their clinical expertise to provide input on developing service assignment exclusions for the measure. In June, the goals of excluding services and the guiding principles, as outline by the TEP, were presented to the workgroup. Acumen's clinicians prepared draft service exclusion rules defined for each Major Diagnostic Category (MDCs), a natural starting point for grouping Medicare Severity-Diagnosis Related Groups (MS-DRGs). The rules were defined using a web-based tool which populated all services and diagnosis combinations from all claims settings representing significant costs during MSPB Clinician episodes. The draft list was summarized into clinically meaningful categories for each MDC to serve as the starting point for discussion with the workgroup. Each MDC and clinical category was systematically discussed to determine which categories of services to exclude. The workgroup also had an opportunity in this meeting to refine the groupings of MDC for which rules would be defined and defined high-level rules applicable to all MS-DRGs (i.e. exclude all hospice service costs). The meeting was followed by a survey in which members could accept or decline the level at which rules are defined and exclusion rules. The service exclusions were iterated based on the discussion and survey results from the first meeting in June 2018 and re-discussed in July 2018 following the same systematic process.

Field Testing

During the feedback period, 5,153 field test reports for the MSPB Clinician measure were downloaded by 436 clinician groups (TINs) and 4,717 clinicians (TIN-NPIs). Stakeholder comments from field testing were summarized for the workgroup and TEP 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.

Data Provided During Field Testing:

Each MSPB Clinician field test report contained the following information:

• The clinician or clinician group's measure score along with the national median score and percentile rank

• Episode cost breakdown by claim type and timing to explain the factors driving the clinician or clinician group measure score (e.g., home health agency, hospice, inpatient, outpatient)

• Episode cost breakdown by categories of service to show the average cost per category (e.g., acute inpatient services, post-acute care)

• Statistics of the TIN or TIN-NPI's specific performance compared to the state and national average (e.g., number of beneficiaries, average standardized cost per beneficiary)

A mock field test report can be viewed on the CMS MACRA Feedback webpage. [1] Along with the Field Test Report, attributed clinicians and clinician groups received an episode-level CSV file that include the risk profile of their attributed episodes.

Education and Outreach:

Acumen directly conducted outreach via email to tens of thousands of stakeholders using the stakeholder contact list developed through previous education and outreach and clinician engagement efforts, as well as CMS, Quality Payment Program, and other available listservs. More detail on this outreach can be found in the Field Test Summary Report on the CMS MACRA Feedback webpage.

Acumen and CMS hosted two office hour sessions in October 2018, 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 50 attendees.

Acumen and CMS hosted a national field testing webinar on October 9, 2018 to provide an overview of the measures being field tested and the information available for public comment. The webinar consisted of an hour-long presentation, outlining (i) the cost measure development activities, (ii) field testing activities, (iii) how to access and understand the confidential field test reports, and (iv) the contents of the reports. The presentation was followed by a 30-minute Q&A session.

A post-field testing webinar was held on March 27, 2019 to provide an update on the measures following field testing. The 60-minute webinar provided an overview of the basics of measure construction, highlighted refinements made after field testing, and provided a summary of testing done on the measures. The presentation was followed by a 30-minute Q&A portion. [2]

Pre-Rulemaking

There was a public comment period after the release of the Measures Under Consideration (MUC) list from December 1, 2018, to December 6, 2018, prior to the MAP Clinician Workgroup Meeting. The MAP Clinician Workgroup met on December 12, 2018 to consider measure specifications and testing updates. In accordance with MAP procedure, these documents were not publicly released but were made available to MAP members. Following the release of the Clinician Workgroup's preliminary recommendation, the report was open for a public comment period from December 21, 2018 to January 10, 2019. The MAP Coordinating Committee met on January 22-23, 2019, to consider these comments alongside the Clinician Workgroup's recommendation. Both MAP meetings were open to the public.

Rulemaking

During the public comment period for the proposed rule from August 14, 2019, to September 27, 2019, stakeholders could review the proposed rule language, measure specifications, and National Summary Data Report when submitting comments. CMS conducted email outreach via its listserv to notify stakeholders about the release of the proposed rule.

[1] CMS, "Revised MSPB Clinician Measure Mock Field Test Report," MACRA Feedback Page, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/Mock-report-for-revised-MSPB-Clinician.pdf.

[2] CMS, MACRA Cost Measures Post-Field Testing Webinar, Quality Payment Program, https://qpp-cm-prod-content.s3.amazonaws.com/uploads/521/MACRA%20Cost%20Measures%20Post%20Field%20Testing%20_Slides.pdf.

U.2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1. Describe how feedback was obtained.

The overarching feedback that was received on measure performance and implementation from the measured entities and others included comments that (i) the revised specifications made several improvements to the MSPB measure; (ii) while field test reports and other supplementary materials were helpful, the complexity of these documents was a challenge to some stakeholders; and (iii) general questions on and proposed updates to the MSPB Clinician measure's attribution methodology. This feedback is detailed in sections U2.2.2 and U2.2.3, with references to publicly-available feedback where appropriate.

TEP and MSPB Service Refinement Workgroup

Input from the MSPB Service Refinement workgroup was gathered via two post-webinar surveys. Fifteen out of 25 members filled out the first survey, and twelve out of 25 members filled out the second survey.

The TEP reconvened in November 2018 to review feedback received on the measure from field testing and discuss any potential refinements to the measure. Ten of the 19 TEP members attended the webinar. Finally, to gather a formal record of the TEP's systematic input and iterative assessments of the measure refinements throughout this process, TEP members completed a face validity survey in November 2019. Fifteen of the 19 TEP members completed the survey.

Field Testing

In total, Acumen received 67 survey responses and 25 comment letters, including many 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.

Pre-Rulemaking

CMS received 14 comments specifically on the MSPB Clinician cost measure included in the Measures Under Consideration List released in December 2018. After the MAP Clinician Workgroup meeting in December 2018, there was another public comment period on their preliminary recommendations, which received six comments specific to the MSPB Clinician cost measure. [1] These public comment periods were facilitated by NQF. Stakeholders were able to submit their comments via the NQF website.

Rulemaking

CMS received over 41,943 comments on the CY 2020 Physician Fee Schedule Proposed rule. A search on the regulations.gov website returns 68 results for "mspb" as a rough approximation of the number of comments on the MSPB Clinician measure during rulemaking. Stakeholders could submit comments through the Federal Register website or via mail.

[1] Measure Applications Partnership, National Quality Forum, http://public.qualityforum.org/MAP/MAP%20Clinician%20Workgroup/2018-2019%20Clinician%20Workgroup%20Archive/MAP_Clinician_Workgroup_Discussion_Guide.html#COMMENTMUC2018-148MIPS.

U.2.2.2. Summarize the feedback obtained from those being measured.

TEP and MSPB Service Refinement Workgroup

During the November 2018 meeting, the TEP agreed with the revisions implemented to the measure prior to field testing and confirmed that they did not believe further refinements were needed. Additionally, Acumen received a generally positive feedback from the TEP members on the MSPB Clinician measure in the face validity survey.

Field Testing

The Field Testing Feedback Summary Report presents all feedback gathered during the field testing period. [1] The following list synthesizes some of the key points that were raised through the field testing feedback period:

• Service exclusion codes and logic developed for the MSPB Clinician episodes make the measure more actionable. Stakeholders expressed appreciation for the development of codes and logic that define a list of services that are unlikely to be influenced by the clinician's care decisions and exclude clinically unrelated services to calculate episode observed cost. In addition to this comment, one stakeholder noted some additional codes for removal from the measure.

• Improved measure better captures clinicians responsible for a beneficiary's health care costs, but can be further refined. One commenter expressed concern that the measure might not reflect the fact that the clinician group practice keeps moderately sick patients out of the inpatient setting and has a disproportionate number of sicker people with inpatient stays. The commenter also indicated that under the current measure calculation methodology this would make the group practice have an unreasonably high MSPB Clinician measure score.

• Field test reports and supplementary materials present useful information for understanding the measure, though reduced complexity could encourage more clinician participation. Stakeholders praised the content of the field test reports and supplementary materials and commented that overall the methodological changes are described in a clear and concise manner. However, the complexity of the information presented in the reports was a challenge for some stakeholders. Specifically, several stakeholders noted that the methodology to attribute MSPB Clinician episodes at the TIN and/or TIN-NPI level was too complex. Some commenters also requested additional information to increase their understanding of the measure.

Pre-Rulemaking

The MAP gives feedback on performance measures from a wide variety of perspectives, with representatives including "consumers, businesses and purchasers, laborers, health plans, clinicians and providers, communities and states, and suppliers." [2] The Clinician Workgroup specifically aims to "ensure the alignment of measures and data sources to reduce duplication and burden, identify the characteristics of an ideal measure set to promote common goals across programs, and implement standardized data elements." [3]

Rulemaking/Public Comment

CMS received comments on the proposed measures during the public comment period for the CY 2020 Physician Fee Schedule proposed rule. Measure-specific comments were received on the measure specifications, which CMS and Acumen review to determine whether changes needed to be made to the measure specifications. For more detailed information on the comments received on the measures as part of the proposed rule public comment period, please see the revised cost measures section in the CY 2020 Physician Fee Schedule final rule for a summary of the public comments received along with CMS responses: https://www.federalregister.gov/documents/2019/11/15/2019-24086/medicare-program-cy-2020-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other.

[1] CMS, Quality Payment Program, "October-November 2018 Field Testing Feedback Summary Report for MACRA Cost Measures," https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/2019-ft-feedback-summary-report.pdf.

[2] National Quality Forum, Measure Applications Partnership https://www.qualityforum.org/Setting_Priorities/Partnership/Measure_Applications_Partnership.aspx.

[3] National Quality Forum, MAP Member Guidebook http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80515.

U.2.2.3. Summarize the feedback obtained from other users.

Pre-Rulemaking

The MAP recognized the importance of cost measures to the MIPS program and conditionally supported the MSPB Clinician cost measure pending NQF endorsement. Specifically, the MAP urged CMS to continue testing the changes to this measure, which are removing costs that are unlikely related to the clinician and a new attribution model, to ensure that they produce the intended results. In particular, MAP noted the need to ensure the measure demonstrates validity and reliability at the National Provider Identifier (NPI) level. MAP also noted the desire to avoid double counting clinician costs in the total cost measures and the episode-based cost measures and for CMS to consider consolidating the MSPB Clinician measure with the Total Per Capita Cost measure also used in MIPS to avoid overlap. MAP suggested that CMS should monitor for unintended consequences to patients such as under treatment, impact on technology innovation, and access to treatment for high-risk, high-resource use patients. Lastly, MAP urged CMS to continuously test and refine the risk adjustment model and incorporate social risk factors, when appropriate.

U.2.3. Describe how the feedback described in 4a2.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not

Field Testing

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 expert clinician workgroup, along with 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.

After careful consideration of field testing analyses and stakeholder feedback, no refinements were made to the measure after field testing . This is because, after reviewing the measure-specific feedback and empirical analyses, the TEP members agreed that the current measure specifications addressed the proposed updates from stakeholders, some of which stemmed from stakeholders misunderstanding the measure methodology. Therefore, the TEP did not propose any additional revisions to the measure.

Rulemaking/Public Comment

During the public comment period for the CY 2020 Physician Fee Schedule proposed rule, stakeholders submitted comments on the proposed all-cost measures, including the MSPB Clinician measure. After receiving public comments, Acumen performed analyses to evaluate the need for and the impact of the proposed updates to the measure.

Given that the proposed updates from the public comments (i) would likely require program-level reporting changes, (ii) stemmed from stakeholder misunderstanding of the current measure specifications, and (iii) were specification updates that would impact the measure intent, the revised MSPB Clinician measure was finalized as proposed.

U.3.1. Progress on Improvement. (Not required for initial endorsement unless available.) Performance results on this measure (current and over time) should be provided in IM.1.2 and IM.1.4.

Discuss:

- **Purpose Progress (trends in performance results)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A.

U.3.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

N/A.

U.4.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

N/A. There were no unexpected findings during the development and testing of this measure.

U.4.2. Please explain any unexpected benefits from implementation of this measure.

N/A. There were no unexpected benefits during the development and testing of this measure.

Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

H.1. Relation to Other NQF-endorsed Measures

If there are related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

H.1.1. List of related or competing measures (selected from NQF-endorsed measures)

2158 : Medicare Spending Per Beneficiary (MSPB) - Hospital

H.1.2. If related or competing measures are not NQF endorsed please indicate measure title and steward.

N/A.

H.2. Harmonization

H.2.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):
Are the measure specifications completely harmonized?

Yes

H.2.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

H.3. Competing Measure(s)

H.3.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

The MSPB Hospital and MSPB Clinician measures are closely aligned. Both measures assess costs from the same time window (three days prior to the index admission to 30 days after discharge) and focus on the same target population of beneficiaries admitted to the inpatient setting. Together, these measures align the incentives for clinicians and hospitals taking care of Medicare patients who are hospitalized.

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services

Co.2 Point of Contact: Ronique, Evans, Ronique.Evans1@cms.hhs.gov, 410-786-3966-

Co.3 Measure Developer if different from Measure Steward: Acumen, LLC

Co.4 Point of Contact: N/A., N/A., macra-cost-measures-info@acumenllc.com, 650-558-8882-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

Technical Expert Panel Members:

Adolph Yates, American Academy of Orthopaedic Surgeons

Alan Lazaroff, American Geriatrics Society

Allison Madson, American Society of Cataract and Refractive Surgery

Alvia Siddiqi, American Academy of Family Physicians

Anupam Jena, Harvard Medical School

Caroll Koscheski, American College of Gastroenterology

Chandy Ellimoottil, American Urological Association

Diane Padden, American Association of Nurse Practitioners

Dyane Tower, American Podiatric Medical Association

Edison A. Machado, Jr., The American Health Quality Association J

ackson Williams, Dialysis Patient Citizens

James Naessens, Mayo Clinic

John Bulger, American Osteopathic Association

Juan Quintana, American Association of Nurse Anesthetists

Kata Kertesz, Center for Medicare Advocacy

Kathleen Blake, American Medical Association

Mary Fran Tracy, National Association of Clinical Nurse Specialists

Parag Parekh, American Society of Cataract and Refractive Surgery

Patrick Coll, University of Connecticut Health Center

Shelly Nash, Adventist Health System

Sophie Shen, Johnson and Johnson Health Care Systems, Inc.

MSBP Service Refinement Workgroup Members:

Adolph Yates, American Association of Hip and Knee Surgeons

Clemens Schirmer, American Association of Neurological Surgeons

Dheeraj Mahajan, AMDA – The Society for Post-Acute and Long-Term Care Medicine

Jennifer Bracey, Society of General Internal Medicine

Steve Sentovich, American Society of Colon and Rectal Surgeons

Marc Raphaelson, American Academy of Neurology

Evan Lipsitz, Society of Vascular Surgery

Caroll Koscheski, American College of Gastroenterology

Richard Dutton, American Society of Anesthesiologists

Kathleen Blake, American Medical Association

Anupam Jena, Harvard Medical School

Alec Koo, American Urological Association

Sanjay Samy, Society of Thoracic Surgeons

Juana Hutchinson-Colas, American Urogynecologic Society

Naakesh Dewan, American Psychiatric Association

Jayme Lieberman, American College of Surgeons

Tracey Weisberg, American Society of Clinical Oncology

William Borden, American College of Cardiology

Anders Chen, Society of General Internal Medicine

Nancy Greenwell, American Academy of Family Physicians

Stephen Lahey, Society of Thoracic Surgeons

Terrry "Lee" Mills, American Academy of Family Physicians

Peter Ray, American Society of Plastic Surgeons

Barbara Spivak, American Medical Association

Robert Lorenz, American Academy of Otolaryngology – Head and Neck Surgery

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision:

Ad.4 What is your frequency for review/update of this measure?

Ad.5 When is the next scheduled review/update for this measure?

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: