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 #: 3575

De.2. Measure Title: Total Per Capita Cost (TPCC)

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

De.3. Brief Description of Measure: The Total Per Capita Cost (TPCC) measure assesses the overall cost of care delivered to a beneficiary with a focus on the primary care they receive from their provider(s). The TPCC measure score is a clinician's average risk-adjusted and specialty-adjusted cost across all beneficiary months attributed to the clinician during a one year performance period.

The measure is attributed to clinicians providing primary care management for the beneficiary, who are identified by their unique Taxpayer Identification Number and National Provider Identifier pair (TIN-NPI) and clinician groups, identified by their TIN number. Clinicians are attributed beneficiaries for one year, beginning from a combination of services indicate that a primary care relationship has begun. The resulting periods of attribution are then measured on a monthly level, assessing all Part A and Part B cost for the beneficiary for those months that occur during the performance period. The beneficiary populations eligible for the TPCC include Medicare beneficiaries enrolled in Medicare Parts A and B during the performance period.

IM.1.1. Developer Rationale: Effective primary care management can support Medicare savings in several ways. For example, more effective primary care management can improve the treatment of chronic conditions by obviating the need for high-cost hospital or emergency department services. It can also direct a greater proportion of patients to lower hospital costs for inpatient services. [1] Given the potential for decreasing spending through improvements in primary care delivery, the TPCC measure allows for a savings opportunity by capturing the broader healthcare costs influenced by primary care.

[1] "Valuation of Care Management Performed by Primary Care Services: An Issue Brief." American Academy of Family Physicians, 2018.

De.1. Measure Type: Cost/Resource Use

S.5. Data Source: Assessment Data

Claims

Enrollment Data

Other

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

New Measure Submission

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.

- The focus of this measure is to assess the overall cost of care delivered to a beneficiary with a focus on the primary care they receive from their provider(s).
- The measure calculates a clinician's average risk-adjusted and specialty-adjusted cost across all beneficiary months attributed to the clinician during a one year performance period. The measure is attributed to clinicians providing primary care management for the beneficiary, who are identified by their unique Taxpayer Identification Number and National Provider Identifier pair (TIN-NPI) and clinician groups, identified by their TIN number.
- The developer states that more effective primary care management can improve the treatment of chronic conditions by avoiding high-cost hospital or emergency department services.
- The developer cites research that shows how primary care management in certain settings, such as Patient-Centered Medical Homes (PCMH), can reduce the total cost of care by reducing utilization of high-cost services.

1b. Opportunity for Improvement.

- The developer provides data demonstrating that Tocal Per Capita Cost (TPCC) has a range of cost performance at the TIN and the TIN NPI levels. Specifically, the interquartile range of performance for TIN level scores is $255 and mean performance of $1,109 for 74,191 clinician group practices. The interquartile range of performance for TIN-NPI is $297 and mean performance of $1,169 for 335,480 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.**
- **yes, this affects all Medicare benes and there is variation. However, the interquartile range is on the order of $250-$275 which seems rather small for trying to weed out variation. Question is what differentiates providers in the 10th and 90th percentile and how comparable are these?**
- **Yes**
- **Developer states that more effective primary care management can improve the treatment of chronic conditions by avoiding high-cost hospital or ED services – would appreciate a citation to a study that shows that effect; developer only cites to research on Patient-Centered Medical Homes, but that is a different model than the full Medicare FFS population and it's hard to hold clinicians to the same standards for cost/resource utilization when the payment is not designed to do so.**
- **Yes.**
- **No concerns**
- **Yes, based on high resource use/cost**

1b. Opportunity for improvement: Was current performance data on the measure provided? Has the

Comments:

- **Fair variation.**
- **No**
- **Yes. The interquartile range of performance for clinician group practice (identified by their TIN) is $255 and mean performance of $1,109 for 74,191. The interquartile range of performance for TIN-NPI is $297 and mean performance of $1,169 for 335,480 practitioners. These cost variations reflected in IQRs indicate room for improvement.**
- **yes, there is a lot of variation. There is a very large standard deviation around the mean which shows a lot of variation (makes me wonder how comparable the scores). The mean ranges from abour $150 to $2700 from the 10th to the 90th percentile.**
- **Yes**

- **if this is looking at all clinicians in FFS and not those in value-based care models it is a challenge to say we're going to hold you to the same standard without the payment design to support it. It is not surprising that the developer brings data that the measures has a range of cost performance at the TIN and the TIN NPI levels considering that clinicians and practices are not uniformly invested in value-based care. Specifically, the interquartile range of performance for TIN level scores is $255 and mean performance of $1,109 for 74,191 clinician group practices. The interquartile range of performance for TIN-NPI is $297 and mean performance of $1,169 for 335,480 practitioners.**
- **The developer provided the IQR for clinicians and clinician group practices. Both were < $300. This calls into question whether there is significant variation across providers.**
- **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-6, L-0, I-0 (Pass)
Methods Panel Individual Validity Ratings: H-1, M-4, L-2, 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 90 – 91.

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 developers used two different measures of reliability: 1) Reliability score (signal to noise) to evaluate the extent to which variation in the measure is due to true, underlying differences in provider performance (signal) rather than random variation (noise) and 2) split-sample reliability testing (intraclass correlation or ICC) to examine agreement between two scores for a facility based on randomly-split, independent subsets of clinician group practice/clinician episodes.
- The performance measure score reliability testing was based on 74,191 clinican groups and 335,480 individual clinicians with 20 or more episodes in the measurement period of 2017-2018.
- The developer reported that the mean reliability score for all clinician group practices was 0.84 with range of 0.77 (25th percentile) to 0.95 (75th percentile). For the 335,480 individual practioners, the mean reliability was slightly higher at 0.88 with range of 0.83 (25th percentile) to 0.95 (75th percentile). When examined by clinician group size, the average reliability score ranged from 0.81 (1 clinician) to 0.94 (21+ clinicians).
- The developer reported that the ICC for the overall sample was 0.76 with 95% confidence interval of 0.75-0.77. The ICC for 68,413 clinican groups as measured by Pearson correlation coefficient was 0.76 and for 265,106 individual practitioners was 0.64.

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *Do you have any concerns with the reliability testing that was not identified by the Scientific Methods Panel?*

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:

- **1.No data element testing. Assumption is that claims data is sufficiently accurate. Consistent approach to data element assessment for CMS claims based measures. NEEDS FURTHER DISCUSSION. 2. Also see discussion in 5.2a above about exclusion of "clinically unrelated services. 3. Standardized pricing has strengths and limitations in understanding resources used.**
- **A provider performance measure of this nature should be documenting deviations from risk-adjusted costs, whereas it appears that this measure simply shows risk-adjusted costs. Also, new patient attibutation to a provider seems like it should be diagnosis based and not based on care provided (e.g. HCPCS/CPT-based).**
- **No additional concerns.**
- **the developer did test retest with split sample and the correlation was moderate. The reliability of the measure with an n=20 was surprisingly high.**
- **The use of beneficiary months (1/13th of the year) instead of calendar months makes sense in creating equivalent spans for assessing cost. However, since "beneficiary month" is also used as the final scalar in the score calculation, it may be misunderstood or misinterpreted by individuals familiar with typical PMPMs or monthly costs.**
- **none**
- **No concerns that were not raised by the Scientific Methods Panel.**
- **No concerns**
- **Specifications are reliable**

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:

- **"SN analysis reaches 0.8 for overall and for all practice sizes except bottom of distribution of 1 clinician practice, where it is 0.73. It should be noted that over half of practices are 1 clinician practices. Split sample analysis does not reach acceptable levels for provider specific comparisons, with TIN 0.76 and TIN-NPI 0.64. Overall reliability probably meets standards for endorsement, notwithstanding the low split sample ICC. However, given low reliability of bottom percentiles of 1 clinician practices (half of practices) and wide distribution of high/low score distributions for TINs and TIN-NPIs at low range of beneficiary months as reported in appendix table 2b.4.2a, I am concerned that the reliability is low for low volume clinicians that meet the threshold of 35 patients."**
- **As healthcare utilization and prices are "real" and costs are merely the sum of price-weighted healthcare utilzation, it is not clear to me how the notion of measure reliability fits here.**
- **None.**
- **The split sample results show moderate correlation, raising questions about repeatablity. yes, the data elements come from administrative data. reliability results show a mean of 0.88 which is good for distinguishing differences across providers**
- **Yes. Based on some of the reliability testing results, I was interested in the rationale for selecting 20 beneficiaries as the required minimum for measurement. I may have missed this in my review of the documentation.**
- **Concern about reliability testing using two years of data when measurement will be over one year – is it actually reliable if one year of data is used? We have concerns with the lack of information on reliability results below the 25th percentile, particularly in light of the statement that CMS generally considers 0.4 to be the threshold for moderate reliability and that 100% of practices and clinicians with at least 20 episodes meet. Some believe that the minimum acceptable thresholds should be 0.7.**
- **No concerns that were not raised by the Scientific Methods Panel.**
- **No concerns**
- **Moderate Reliability**

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:

- Attribution:
  - This measure is attributed to clinicians and clinician groups. This attribution approach was developed in order to encourage providers to facilitate care coordination, assess referral practices, and support their role in improving costs and resource use.
- Costing approach:
  - The costing approach is based on payments by Medicare for services within the identified resource use service categories. Payments are based on agreed upon fee schedules for each setting.

2b2. Validity Testing:

- The developer conducted face validity testing with an expert panel. The developer reported that 80% (12 out of 15) of the experts agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost effectiveness
- The developer also conducted dmpirical validity testing with known indicators of resource or service utilization, complications related to acute admission, and post-acute care utilization. Four clinical themes were created around inpatient service, post-acute care (PAC), emergency services not included in an admission and outpatient E&M services, procedures, and therapy.
- The developer reported that the mean of beneficiary's average risk- and specialty-adjusted monthly cost for a beneficiary during the measurement period is $1,187. The mean of beneficiary's average risk- and specialty-adjusted monthly cost for beneficiaries with services relating to acute inpatient admissions is $2,647, compared with $866 for a beneficiary without acute inpatient admissions. The mean of beneficiary's average risk- and specialty-adjusted monthly cost with services relating to Post-Acute Care is $2,427 compared with $996 for a beneficiary without PAC.
- The results from the clinical themes analysis found the correlation with risk- and specialty-adjusted cost were low to moderate. At both the TIN and TIN-NPI levels, there is a moderate correlation between the Skilled Nursing Facility service category and risk-adjusted cost (0.54), low correlation between Outpatient E&M Services, Procedures, and Therapy and risk-adjusted cost (0.45) and Acute Inpatient Services (0.38), and very low for the HH category (0.11), Non-Hospital Admission Emergency Services (0.15).

Clinical Theme	Pearson Correlation	
	TIN	TIN NPI

Acute Inpatient Services	0.38	0.38
Emergency Services Not Included in Hospital Admission	0.15	0.15
Outpatient E&M Services, Procedures, and Therapy	0.45	0.45
Post-Acute Care: Home Health	0.11	0.11
Post-Acute Care: IRF/LTCH	0.18	0.18
Post-Acute Care: SNF	0.54	0.54

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

- The developer reports 15.3% of episodes were excluded because of one or more exclusion criteria

2b4/2c. Risk adjustment

- The developer uses CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) models for new enrollees, continuing enrollees, enrollees in long-term institutional settings, and uses CMS End Stage Renal Disease Version 21 (CMS-ESRD V21) models are used for new enrollees with ESRD, and community enrollees with ESRD.
- The developer reported results showing that dual enrollment is associated with systematically higher cost. The addition of AHRQ SES index was significant and negative in value for the community, institutional, and new enrollee models, but not found to be significant in either the dialysis or new enrollee dialysis models. The developer reported that inclusion of SES in the model did not significantly change TIN or TIN/NPI performance scores on average.
- The developer reported that the R-squared for the CMS-HCC V22 model for community enrollees, segmented by dual eligibility and disability, ranged from 0.11 to 0.12. The CMS-ESRD v21 R-squared values are 0.02 and 0.11 for the dialysis new enrollee and dialysis community models, respectively.

2b5: Meaningful Differences

- The developer reported that 16.8% of TIN's and 10.9% of TIN-NPI's had scores that were significantly lower than the mean while 17.9% of TIN's and 11.4% of TIN-NPI's had scores that were significantly higher than the mean.

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *Does the Standing Committee have any concerns regarding this amount of exclusions?*
- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *Do you agree with the developer's rationale for not included SES factors in the risk-adjustment model?*

Guidance from validity 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 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:

- **Validity is assessed by: Face validity. Panel of experts. Empirical testing. Comparisons of distribution of costs across percentiles for patients with and without inpatient treatment and with and without post acute care. Correlation of predicted costs and actual costs at TIN and TIN-NPI level for acute inpatient care, Emergency services, Outpatient E&M, PAC-Home Health, PAC-IRF/LTCH, and PAC-SNF. Attribution to clinician validity tested based on percentage of primary care E&M provided by attributed TIN or TIN-NPI. : The key question on face validity for me was ""The scores obtained from the TPCC measure as specified will provide an accurate reflection of the costs for overall primary care, and can be used to distinguish good and poor performance on cost effectiveness." 12 of 15 agreed with an average score on a 6 point scale of 4.8. Major concerns reported for those disagreeing were double counting with episode based care, and attribution issues. I will accept the TEP assessment of the face validity. Correlation of predicted and actual costs and comparison of distribution across patients provide some evidence of validity, but are subject to concern about risk adjustment discussed below. Attribution percentages are at levels accepted in earlier measures.**
- **It seems that costs are face-valid with respect to healthcare utilization. Not sure what "accurate reflection of cost effectiveness" means on page 5. The purpose of the "attribution" method is not sufficiently described, at least that I can find, so it is unclear whether the described approach is suitable for this purpose.**
- **None.**
- **Concern about not including social risk factors**
- **No concerns.**
- **none**
- **No concerns not raised by the Scientific Methods Panel.**
- **No concerns**
- **No concerns**

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

Comments:

- **None**

- **Measuring providers using what appears to be risk-adjusted costs seems to lose the uniqueness of each provider that should be measured. However, the risk-adjustment algorithm on page 11 is insufficiently described, so I may be misinterpreting the measure.**
- **None.**
- **there are differences noted, but I'm really struck by the size of the standard deviation**
- **No concerns.**
- **none**
- **The scientific methods panel reports rather poor r-squared results for the model. In addition, I would be concerned that within specialties, there may be variation in the types of care provider and types of patients a specialist sees that would make their costs not comparable. I would expect such a broad model to be more accurate for certain specialties and patient populations.**
- **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:

- **No specific concerns**
- **A primary care provider is often responsible for referalls and initiating down-stream healthcare utilization. This measure seems to miss this as much as I can tell.**
- **None.**
- **none**
- **It seems like the exclusion of Part C and D costs (as well as other third party payers) makes sense for comparability and feasibility purposes, however, the name of the measure and description do not make this clear.**
- **no**
- **None.**
- **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:

- **Attribution is standard issue in measures like this.**
- **Unclear how referrals are handled**
- **None.**
- **the attribution approach is clear, but given how much of the cost is driven by post acute care, it raises the question of why single accountability for this measure. feels to me like CMS is missing out on oppoprtunity for better coordination through joint accountabilities.**
- **No concerns.**
- **Validity of attribution methodology and whether it is fair. For example, if a patient is receiving outpatient surgery, the surgery center will often require the patient to be cleared for surgery by the PCP. The PCP will likely comply and order an EKG or other necessary tests, and with that be accountable for everything that happens to the patient for the next 12 months. This could disincentivize PCP engagement in the pre-surgery process. On the flipside, a patient who is healthy or whose health problems are appropriately managed and do not return to the PCP for 6 months or a year will not be attributed to the clinician, and those lower costs will not be reflected in the clinician's TPCC average.**
- **I think the ability for a provider to control costs and resources will vary greatly by 1) the other providers a patient is seeing, and 2) the role of the provider in patient care. The model seems to very liberally assign "primary care" responsibilities to providers.**
- **No concerns**
- **No concerns**

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:

- **N/A**
- **None**
- **None.**
- **none**

- **Winsorizing at the 99th percentile makes conceptual sense. From an accountability perspective, it would be helpful to understand this from a cost perspective. Typically, it seems like truncation of annual costs falls somewhere between $50,000 and $250,000 for clinician groups.**
- **SMP panel review seemed concerned with complexity of exclusions -might want to hear more from developer in response to this.**
- **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:

- **The risk adjustment model only includes information on medical history and status at start of assessment period. This has two problems I believe merit discussion: First, for new enrollees, even those with ESRD, the risk adjustment model, which includes only information on Medicaid status, disability status, gender and age, has an R-square of 0.017-0.021. This contrasts with the R-square for the risk adjustment models that include medical history information of 0.114-0.137, not great but in the range we have previously observed for the cost models. The low R-square for new enrollees implies virtually no information about variability of costs. In practices with large numbers of beneficiaries and beneficiary months or few new enrollees, this may not be a problem. For smaller practices, there are increased likelihoods that random health shocks, negative or positive, will substantially affect the Observed to Expected ratio (O/E) for the practice due to changes in the medical condition of a few patients outside of the control of the clinician that require expensive treatment (e.g., newly diagnosed cancer) or due to preexisting conditions a newly enrolled patient brings to a practice (e.g., congestive heart failure). Second, the problem described above may also affect the adjustment of risk for patients who have a one-year look back period. While the O/E ratio is estimated monthly for each patient and averaged over the twelve month period for the measure, if I understand the methods being used, the medical condition and thus the risk model is based on the information from the look back year and not adjusted for new diagnoses during the year. If this is correct, then a TIN or TIN-NPI who is treating a patient who develops or is diagnosed with a new, potentially expensive to treat, condition during the year does not have their expected adjusted. This may be a reasonable approach for entities with large numbers of beneficiaries, such as Part C Medicare Advantage plans or large ACO, since the risk is spread over a reasonable number of patients, but it may not be reasonable for a TIN or TIN-NPI that has a small number of beneficiary months. There is some evidence that this problem exists in Table 2b4.2a, which reports the measure score by clinician characteristics.What one sees looking down the rows from low volume to high volume is that the variability is higher for low volume practices. For TINs, it varies from an average of 533 to 964 at the 10th percentile and 1608 to 1317 at the 90th percentile. At the 99th percentile it varies form 2684 to 1661, and the variance across is volume is even larger at the TIN-NPI level.My concern is that random shocks of either poor health or low volume of use are more significant in low volume practices, and the 12 month look back risk adjuster doesn't protect these small volume practices from having high costs due to new diagnoses attributed to them as excess spending relative to their peers. Either the risk adjustment needs to include dynamic adjustment for new diagnoses or the threshold for inclusion in the measure should be increased. The included data shows correlations of predicted costs with high cost services such as hospitalizations and institutional PAC services. I am concerned the adjustments are not enough. The report cited in the testing document on the analysis of the calibration of the risk adjuster reports lower than predicted costs through the bottom eight deciles and higher than predicted costs in the top two deciles. The explanation for this provided is "This is because the model is not intended to predict the random costs that result in either high costs or low costs in a given year." Larger volumes will tamp this problem down. The question for the committee is whether should be a concern in endorsement for small patient volumes, or whether we should ask for a dynamic model for adjusting expected costs in light of new diagnoses.**

- **Not clear why specialty-adjustment is needed here.**

- **I agree with the SMP member # 3 that SES characteristics should have been included in the final risk-adjustment model. Also, the fact that only 52% and 45% of the E & M claims seem to be billed by the attributed TINs or TIN-NPIs needs discussion.**

- **Concern about not adjusting for social risk factors when assigning responsibility at the NPI/TIN level. This may lead providers to avoid these types of patients. So while no difference on average, for those providers who disproportionately care for patients with social risk factors, not adjusting for within provider differences in social risk factors (e.g., dual status) is problematic.**

- **I'm not an expert here, so I would defer to others, but the r-squared struck me as a bit low (even for a prospective risk adjustment model).**
- **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, which was significant in the community, institutional, and new enrollee HCC models, but not for dialysis or new enrollee dialysis models. It is perplexing to think gender and dual eligibility alone are sufficient. 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 I the model. It does not appear to have been addressed, just seems to be HCC model is enough and workable.**
- **No additional comments.**
- **No concerns**
- **No concerns**

Combined Scientific Methods Panel Preliminary Analysis of Scientific Acceptability

Measure Number: 3575

Measure Title: Total Per Capita Cost (TPCC)

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:**
Panel Member #1: Long-term Minimum Data Set, and Common Medicare Environment
Panel Member #2: Minimum Data Set **(MDS)**

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 Total Per Capita Cost (TPCC) measure assesses the overall cost of care delivered to a beneficiary with a focus on the primary care they receive from their provider(s). The TPCC measure score is a clinician's average risk-adjusted and specialty-adjusted cost across all beneficiary months attributed to the clinician during a one year performance period.

The measure is attributed to clinicians providing primary care management for the beneficiary, who are identified by their unique Taxpayer Identification Number and National Provider Identifier pair (TIN-NPI) and clinician groups, identified by their TIN number. Clinicians are attributed beneficiaries for one year, beginning from a combination of services indicate that a primary care relationship has begun. The resulting periods of attribution are then measured on a monthly level, assessing all Part A and Part B cost for the beneficiary for those months that occur during the performance period. The beneficiary populations eligible for the TPCC include Medicare beneficiaries enrolled in Medicare Parts A and B during the performance period.

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: Identify Beneficiaries for Attribution (i.e., Candidate Events)

A 'candidate event' indicates the start of a primary care relationship between a clinician and beneficiary, and is identified by the occurrence of two Part B Physician/Supplier (Carrier) claims with particular CPT/HCPCS services billed in close proximity. There are two different sets of CPT/HCPCS codes used: E&M primary care services and primary care services.

E&M primary care services are a specific set of evaluation and management codes for clinician visits in the outpatient setting, physician office, nursing facility, or assisted living.

Primary care services are a broader list of services related to routine primary care and generally fall into the following categories: Durable Medical Equipment (DME) and Supplies, Electrocardiogram, Laboratory - Chemistry and Hematology, Other Diagnostic Procedures (Interview, Evaluation, and Consultation), Other Diagnostic Radiology and Related Techniques, Prophylactic Vaccinations and Inoculations, Routine Chest X-ray, Clinical Labs, Preventive Services.

To identify a candidate event, firstly, an initial E&M primary care service billed on Part B Physician/Supplier (Carrier) claim is identified. This E&M primary care service is not considered if it occurs during a beneficiary's stay at a Critical Access Hospital (CAH), Inpatient Facility, or Skilled Nursing Facility (SNF). Secondly, in addition to the initial E&M primary care service, the presence of at least one of the following services confirms the candidate event:

•From any TIN within +/- 3 days: Another primary care service,

•From the same TIN within + 90 days: A second E&M primary care service OR another primary care service

See the "Prim_Care_E&Ms" and the "Prim_Care_Services" tabs of the TPCC Measure Codes List file for the list of the Current Procedural Terminology/Healthcare Common Procedure Coding System (CPT/HCPCS) codes that identify E&M primary care services and primary care services, respectively. The URL of this downloadable file is linked in Section S.1.

STEP 2: Exclude Clinicians Unlikely to be Providing Primary Care

Once candidate events are identified, individual clinicians (identified by TIN-NPI) can be attributed based on their involvement in the candidate event and how their practice relates to primary care. The TIN-NPI assigned responsible for a candidate event is the clinician found on the initial E&M primary care service claim of the candidate event. TIN-NPIs are excluded from attribution if they meet one of two types of exclusions: service category exclusions and specialty exclusions. Candidate events belonging to TIN-NPIs who meet any of these exclusions are removed from attribution and measure calculation for both the TIN-NPI and their respective clinician group (identified by TIN).

STEP 2.1: Exclude Clinicians Based on Service Category Exclusions

Clinicians whose billing patterns indicate that they tend to provide services that are not within the scope of primary care are excluded from attribution of the TPCC measure. A TIN-NPI and all their candidate events are removed from attribution if he or she bills the volume of services below within +/-180 days of a candidate event for a beneficiary:

•At least 15 percent of the clinician's candidate events are billed with 10-day or 90-day global surgery services.

•At least 5 percent of the clinician's candidate events are billed with anesthesia services.

•At least 5 percent of the clinician's candidate events are billed with therapeutic radiation services.

•At least 10 percent of the clinician's candidate events are billed with chemotherapy services.

The list of CPT/HCPCS codes used for each of the service exclusions can be found in the tabs of the TPCC Measure Codes List file labeled: "HCPCS_Surgery," "HCPCS_Anesthesia," "HCPCS_Ther_Rad," and "HCPCS_Chemo." The downloadable file is linked in Section S.1.

STEP 2.2: Exclude Clinicians Based on Specialty Exclusions

Clinicians who – based on their specialty – would not reasonably be responsible for providing primary care are excluded from attribution of the TPCC measure. This exclusion aims to keep primary care specialists and internal medicine subspecialists who frequently manage patients with chronic conditions falling in their areas of specialty. The excluded specialties list contains 56 specialties that fall into the following broad categories:

•Surgical sub-specialties

•Non-physicians without chronic management of significant medical conditions

•Internal medicine sub-specialties with additional highly procedural sub-specialization

•Internal medicine specialties that practice primarily inpatient care without chronic care management

•Pediatricians who do not typically practice adult medicine

The list of HCFA Specialty codes that identify clinicians that are included or excluded from the measure attribution can be found in the "Eligible_Clinicians" tab of the TPCC Measure Codes List. The downloadable file is linked in Section S.1.

STEP 3: Construct Risk Windows

Candidate events that are not excluded initiate the opening of a risk window, a year-long period that begins on the date of the initial E&M primary care service of the candidate event. The performance period is divided into 13 four-week blocks called beneficiary months. Beneficiary months during the risk window are considered attributable if they occur during the performance period. In the event that a risk window begins or ends with a partially covered month, only the portion during the risk window and the performance period is considered for attribution.

STEP 4: Attribute Beneficiary Months to TINs and TIN-NPIs

Beneficiary months are attributed to a TIN according to the following steps:

•Identify the TIN billing the initial E&M primary care service claim of each candidate event.

•Determine beneficiary months that fall within the risk windows of the candidate events that were initiated by the TIN and overlap the performance period and attribute those beneficiary months to the TIN.

Beneficiary months are attributed to a TIN-NPI according to the following steps:

•Identify the TIN-NPI billing the initial E&M primary care service claim of each candidate event.

•Determine beneficiary months that fall within the risk windows of the candidate events that were initiated by the TIN-NPI and that overlap the performance period.

•Identify the TIN-NPI within an attributed TIN that is responsible for the plurality of candidate events provided to the beneficiary. If two or more TIN-NPIs under a single TIN provide the same proportion of candidate events to a beneficiary, attribute the beneficiary to the TIN-NPI that provided the earliest candidate event.

•Attribute only the beneficiary months from candidate events that the TIN-NPI is responsible for initiating, which is not necessarily all candidate events attributed to the TIN for that beneficiary.

STEP 5: Calculate Payment-Standardized Monthly Observed Costs

The monthly observed cost for attributed beneficiary months is the sum of all service costs billed for a particular beneficiary during a beneficiary month. Monthly observed costs are standardized to account for differences in Medicare payments for the same service(s) across Medicare providers. Payment standardization accounts for differences in Medicare payment unrelated to the care provided, such as those from payment adjustments supporting larger Medicare program goals (e.g. indirect medical education add-on payments) or variation in regional healthcare expenses as measured by hospital wage indexes and geographic price cost indexes (GPCIs). Standardized costs that occur during partially covered months are pro-rated, based on the portion of the month covered by the risk window.

STEP 6: Risk-Adjust Monthly Costs

Beneficiary cost may differ across clinicians for reasons unrelated to the attributed clinicians' treatment and outside of their control. Risk adjustment accounts for case-mix of patients and other non-clinical characteristics that influence complexity of case-mix and is defined by a patient's claims found one year prior the start of a respective beneficiary month. The CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) 2016 Risk Adjustment models are used for beneficiaries without End Stage Renal Disease (ESRD). Specifically,

•The new enrollee model is used for beneficiaries that have fewer than 12 months of Medicare medical history. The model accounts for each beneficiary's age, sex, disability status, original reason for Medicare entitlement (age or disability), and Medicaid eligibility.

•The community model is used for beneficiaries that have least 12 months of Medicare medical history. The model includes the same demographic information as the new enrollee model but also accounts for clinical conditions as measured by HCCs.

•The institutional model is used for beneficiaries who were in long-term institutional settings. The model includes demographic variables, clinical conditions as measured by HCCs, and various interaction terms.

The CMS-ESRD Version 21 (CMS-ESRD V21) 2016 Risk Adjustment models are used for ESRD beneficiaries receiving dialysis. Specifically,

•The dialysis new enrollee model is used for ESRD beneficiaries that have fewer than 12 months of Medicare medical history. The model accounts for each beneficiary's age, sex, disability status, original reason for Medicare entitlement (age or disability), Medicaid eligibility, and ESRD.

•The dialysis community model is used for ESRD beneficiaries that have at least 12 months of Medicare medical history. The model includes the same demographic information as the new enrollee model but also accounts for clinical conditions as measured by HCCs.

The "HCC_Risk_Adjust" tab of the Measure Codes List file lists all variables included in the CMS-ESRD V21 and the CMS-HCC V22 risk adjustment models. The downloadable file is linked in Section S.1.

The standardized risk scores from the CMS-ESRD V21 and CMS-HCC V22 models are generated for each beneficiary's month that summarizes the beneficiary's expected cost of care relative to other beneficiaries. Risk scores for ESRD beneficiaries are normalized to be on a comparable scale with the HCC V22 risk scores. A risk score equal to 1 indicates risk associated with expenditures for the average beneficiary nationwide. A risk score greater than 1 indicates above average risk, while a risk score less than 1 indicates below average risk.

The risk-adjusted monthly cost for each attributed month is calculated according to the following steps:

•Calculate CMS risk score for each beneficiary month using diagnostic data from the year prior to the month. This risk score is normalized by dividing by the average risk score for all beneficiary months.

•Divide observed costs for each beneficiary month by the normalized risk score to obtain risk-adjusted monthly costs.

•Winsorize risk-adjusted monthly costs at the 99th percentile by assigning the 99th percentile of monthly costs to all attributable beneficiary months with costs above the 99th percentile.

•Normalize monthly costs to account for differences in expected costs based on the number of clinician groups to which a beneficiary is attributed in a given month. The normalization factor is the inverse cube root of the number of attributed clinician groups for that beneficiary month.

STEP 7: Specialty-Adjust Monthly Cost

The specialty adjustment for the TPCC measure is a cost adjustment applied to account for the fact that costs vary across specialties and across TINs with varying specialty compositions. The specialty adjustment at the TIN and TIN-NPI levels is calculated as follows:

1) Calculate the average risk-adjusted monthly cost for each TIN and TIN-NPI by averaging risk-adjusted monthly cost across all attributed beneficiary months.

2) Calculate the national specialty-specific expected cost for each specialty as the weighted average of TIN/TIN-NPI's risk-adjusted monthly cost.

2a) Define the weight for each TIN/TIN-NPI as the percentage of clinicians with that specialty multiplied by the total number of beneficiary months attributed to the TIN/TIN-NPI multiplied by the number of clinicians with that specialty.

2b) There will only be one specialty designation for a TIN-NPI. Therefore, the percentage of clinicians with a specialty and number of clinicians with a specialty will always be equal to 1.

3) Calculate the specialty-adjustment factor for each TIN or TIN-NPI as follows:

3a) Multiply the national specialty-specific expected cost for each specialty by the respective specialty's share of Part B payment within a TIN or TIN-NPI.

3b) Sum the weighted share of national specialty-specific expected cost calculated in the previous step across all the specialties under a given TIN or TIN-NPI.

STEP 8: Calculate the TPCC Measure Score

Calculate final risk-adjusted, specialty-adjusted cost measure by dividing each TIN and TIN-NPI's average risk-adjusted monthly cost by their specialty-adjustment factor and multiply this ratio by the average non-risk-adjusted, winsorized observed cost across the total population of attributed beneficiary months.

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

Panel Member #1: None

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: No information provided for Measure Importance, Feasibility, or Usability and Use.

Computation methodology is very complex with many moving parts; difficult to follow the logic—and this could especially be true for measured units trying to understand what they can do to improve their score.

Panel Member #7: No concerns but a complicated step by step calculation methodology

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**

   Panel Member #3: X—Not applicable

   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 were conducted at the clinician (TNI) and clinician-group (TNI-NPI), which were the measure entities for this measure, and thus considered appropriate for this measure.

   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 (20 or more beneficiaries 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 approach is acceptable.

Panel Member #7: Split sample reliability testing was utilized over two years of data.

Panel Member #8: ICC/Split sample testing

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: Estimated reliabitlity for TIN was 0.84 and and TIN-NPI was 0.88, using a volume threshold of 20 beneficiaries. It was also reported that 100 percent of TINs and TIN-NPIs at the reporting case minimum have reliability >= 0.4. As opposed to the measure developer's claim that the CMS generally considers 0.4 as the threshold for "moderate" reliability, it is actually the lower limit of the moderate reliability. Interestingly, when the measure score is assessed by practice size, the average reliability scores monotonically increases from 0.81 (1 clinician) to 0.94 (21+ clinicians).

Reliability measaure through split-sample reliability testing though is 0.76 and 0.64 for TNIs and TNI-NPIs, respectively, and thus may be considered moderate.

Panel Member #2: Overall, reliability testing using 74,191 clinician group practices from 2018 indicated good reliability, regardless of clinician group size. The average reliability score for all clinician group practices was 0.84 with range of 0.77 (25^{th} percentile) to 0.95 (75^{th} percentile). For the 335,480 individual practioners, the mean reliability was slightly higher at 0.88 with range of 0.83 (25^{th} percentile) to 0.95 (75^{th} percentile). When examined by clinician group size, the average reliability score ranged from 0.81 (1 clinician) to 0.94 (21+ clinicians). The ICC for the overall sample was 0.76 with 95% confidence interval of 0.75-0.77.

The ICC for 68,413 clinican groups as measured by Pearson correlation coefficient was 0.76 and for 265,106 individual practitioners was 0.64. This shows lower reliability than the signal to noise measure, but still within moderate reliability ranges.

Panel Member #3: The following table provides evidence of reliability given the Reliability score values:

Table 1. Distribution of Reliability Score Results for TINs and TIN-NPIs with an Overall Testing Volume Threshold of 20 Beneficiaries

Reporting Level	Number of TINs or TIN NPIs	Mean (Std. Dev.)	25^{th} Pct.	50^{th} Pct.	75^{th} Pct.
TIN	74,191	0.84 (0.14)	0.77	0.89	0.95
TIN-NPI	335,480	0.88 (0.08)	0.83	0.91	0.95

* Pct. = percentile.

The following table provides evidence of reliability given the split sample ICC score values:

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	**74,191**	**0.84 (0.14)**	**0.77**	**0.89**	**0.95**
1 Clinician	37,657	0.81 (0.13)	0.73	0.85	0.91
2-4 Clinicians	17,042	0.86 (0.13)	0.80	0.91	0.95
5-20 Clinicians	12,842	0.89 (0.13)	0.85	0.94	0.97
21+ Clinicians	6,650	0.94 (0.11)	0.93	0.98	0.99

* Pct. = percentile.

Table 3. Split-sample Intraclass Correlation Coefficients

Reporting Level	# of TINs or TIN NPIs	Mean Score FY 2017	Mean Score CY 2018	Pearson Correlation Coefficient	ICC(2,1)
TIN	68,413	$1,089	$1,089	0.76	0.76
TIN-NPI	265,106	$1,143	$1,143	0.64	0.64

Panel Member #7: Reliability scores showed a mean of 0.84 with a standard deviation of 0.14 for TIN and a mean of 0.88 with a standard deviation of 0.08 for TIN-NPI. The mean reliability increased from 0.81 to 0.94 with clinician group size. Pearson correlation coefficient was 0.76 for TIN's and 0.64 for TIN-NPI.

Panel Member #8: No concerns.

Panel Member #9: Resutls 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 my rationale in 7 above

    Panel Member #2: Reliability testing using two different approaches shows moderate to high reliability, but reliability was lower based on ICC. Showed variability in reliability scores based on practice size (# of clinicians) which was good, but DID NOT show variability in reliability scores at the individual practitioner level based on # of beneficiaries attributed to the practitioner. Would like to see this.

    Panel Member #3: Results seem positive assuming that the measure is appropriate.

    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: The data supported a good measure score reliability

    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 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: Again, the measure is quite complex. I may be not a good judge of whether the exclusions are appropriate.

    Panel Member #7: Not continuous enrollment in Medicare Part A and B or any enrollment in part C contributed the largest number of exclusions and with outside residency and railroad workers resulted in almost 15.3% exclusions. Analysis demonstrated a variation in the observed costs in these populations.

    Panel Member #8: No concerns.

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: No concerns – The measure as demonstrated in section 2b4 of the testing document will capture meaningful differences in provider performance.

    Panel Member #2: NONE

    Panel Member #7: 16.8% of TIN's and 10.9% of TIN-NPI's had scores that were significantly lower than the mean while 17.9% of TIN's and 11.4% of TIN-NPI's had scores that were significantly higher than the mean.

    Panel Member #8: No concerns.

14. **Panel Member #9**: No concern.**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 #8: NA
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 measure developer reports the number of providers (TINs or TIN/NPIs) with at least one beneficiary being excluded due to one of the exclusion criteria. To me, that's not missing data – those patients are justifiably excluded due to not meeting the measure criteria (or, satisfy the exclusion restrictions).

Panel Member #2: NONE

Panel Member #7: Missing data is provided above for the non-continuous enrollment, outside residency, and railroad workers populations.

Panel Member #8: NA

Panel Member #9: No concern.

16. **Risk Adjustment**

16a. **Risk-adjustment method ☐ None ☒ Statistical model ☒ Stratification**

Panel Member #2: Stratification by 5 risk categories

Panel Member #3: Statistical model (28-133 RFs)**.** Stratification (5 risk categories)

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

Panel Member #3: ☒ —Not applicable (NOTE—if you change the question to "For factors not present…", then a simple Yes/No works.)

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

Panel Member #2: Yes applies to clinical risk factors, NO SES factors were included.

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

Panel Member #1: In order to account for differences in patient case-mix, CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) models were used for new enrollees, continuing enrollees, enrollees in long-term institutional settings, while for beneficiaries with ESRD, the CMS ESRD Version 21 (CMS-ESRD V21) models are used for new enrollees with ESRD, and community enrollees with ESRD. Although

originally developed and tested for Medicare Advantage population, the measure developer demonstrated why and how these models can be applied to the traditional Medicare fee-for-service populations as well.

Table 9 of the testing document shows the range in mean risk- and specialty-adjusted monthly cost across deciles to be $1,013 to $1,396. However, I would have expected the costs to be monotonically increasing by risk score with higher risk patients having higher costs; however, The numbers in the Table 9 are somewhat counter-intuitive to me as the patients in the lower risk deciles had higher costs than those with median risk!

Panel Member #2: The developers present a strong conceptual argument for including SES and statistical results show significant results based on p-values indicating that SES factors are predictive for determining beneficiary cost. The T-tests revealed significant p-values for the majority of factors that interact with gender or dual, indicating that social risk factors are predictive of resource use among beneficiaries for the relevant characteristic. For the community, institutional, and dialysis models, dual enrollment is associated with systematically higher cost. The addition of AHRQ SES index was significant and negative in value for the community, institutional, and new enrollee models, but not found to be significant in either the dialysis or new enrollee dialysis models. They found inclusion of SES in the model did not significantly change TIN or TIN/NPI performance scores on average. They found high correlation between measure costs based on risk adjustment models and SES factors included.

In spite of finding SES was significant predictor in the models, they chose not to include them in final model. They did not share differences in mean cost per beneficiary based on presence of selected SES. I do not agree with their rationale to only include gender and dual status as proxies for SES as that is what CMS includes in HCC payment model based on significance of SES as predictors of cost.

The models as specified have fair to good discrimination properties based on clinical risk adjustments applied. The R-squared for the CMS-HCC V22 model for community enrollees, segmented by dual eligibility and disability, range from 0.11 to 0.12. The CMS-ESRD v21 R-squared values are 0.02 and 0.11 for the dialysis new enrollee and dialysis community models, respectively.

Panel Member #3: The R-squared results are rather poor. The R-squared reported in the December 2018 CMS Report to Congress for the CMS-HCC V22 model for community enrollees, segmented by dual eligibility and disability, range from 0.11 to 0.12. The CMS-ESRD v21 R-squared values are 0.02 and 0.11 for the dialysis new enrollee and dialysis community models, respectively.

Panel Member #7: The extensively utilized CMS-HCC V22 and CMS-ESRD V21 were utilized for risk adjustors.

Panel Member #9:Risk-adjustment approach is acceptable.

<u>For cost/resource use measures ONLY</u>:

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

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

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

    Panel Member #1: With regard to attribution, the mean share of beneficiary's E&M claims billed by attributed TINs or TIN-NPI is 52.8 percent and 45.0 percent, respectively, which seem small to me. This would mean that 47% and 55% of beneficiary's E&M claims will be billed by non-attributed TINs and TIN-NPI. Doesn't it suggest that the evaluation of performance of the attributed TINs and TIN-NPI may be based only on (approximately) half or lower fraction of the beneficiaries that they are attributed to?

    Although not specifically stated, I assumed that winsorizing accounted for possible outliers in the cost data.

Panel Member #3: Validity evidence is thin at best.

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: Evaluation of face validity for the TPCC measure was conducted via a structured process for gathering detailed input from experts and a broad range of stakeholders. The entire process of face validity is described very well in the testing document, and I have no concerns.

    Empirical validity of the TPCC measure was captured two ways: first, confirm that the TPCC measure captures variation in service utilization by examining differences in mean risk- and specialty-adjusted cost for beneficiary months stratified by beneficiaries with known indicators of resource or service utilization, specifically complications related to acute admission and post-acute care utilization. Second, conduct empirical testing as to whether the measure is capturing variation in provider cost in the manner intended by evaluating how different types of cost within a clinical theme (e.g., acute inpatient service, post-acute care) impact risk- and specialty-adjusted monthly costs.

    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 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 mean risk- and specialty-adjusted monthly cost for beneficiaries with and without complications related to acute admission and post-acute care utilization occurring in the measurement period. They hypothesized that beneficiaries as measured by TPCC with these indicators of resource or service utilization would be more expensive those without.

    2. Empirically tested whether the measure is capturing variation in provider cost by evaluating how different types of cost impact risk- and speciality-adjusted measure scores. Certain services or costs included in the TPCC 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 in each service category and the risk- and speciality-adjusted cost. They also examined the possibility of testing a hypothesized relationship between clinicians' TPCC scores and their scores on MIPS quality measures, but determined they were unable to do this testing.

Finally, they conducted two types of validation testing on their method of attribution of patients to clinicians using E&M codes and also an impact analysis on volume of TINS attributed to the measure soley baed on services conducted by nurse practitioners and/or physician assistants.

Panel Member #3: Three types: Face, Empirical validity, and Attribution.

Panel Member #5: ??

Panel Member #7: Face validity with an expert panel was followed by empirical validity testing with known indicators of resource or service utilization, complications related to acute admission and post-acute care utilization. Also, four clinical themes were created around inpatient service, post-acute care, emergency services not included in an admission and outpatient E&M services, procedures, and therapy.

Panel Member #8:

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: With regard ot assessment of face validity, 80% (12 out of 15) of the experts agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost effectiveness.

With regard the first approach of estblishsing empirical validity, the average risk- and specialty-adjusted monthly costs for beneficiaries with acute inpatient admissions and post-acute care in the measurement period are higher than for beneficiaries without those services implying that the TPCC measure can capture higher resource utilization by clinicians who have higher rates of complications related to acute inpatient services and post acute care, while not disincentivizing the provision of appropriate care in other areas (e.g., outpatient services).

The second approach of clinical theme analysis was less convincing. The correlation between the SNF service category and risk-adjusted cost was only 0.54 at both the TIN and TIN-NPI levels, which was dubbed as a strong correlation by the measure developer. However, this is perhaps only moderate correlation. All the correlation estimates across other clinical thems were less than 0.54.

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, 12 (80%) agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost effectiveness, indicating good face validity.

Results of empirical validity testing found the average risk- and specialty-adjusted monthly costs for beneficiaries with acute inpatient admissions and post-acute care in the measurement period are higher than for beneficiaries without those services. The mean of beneficiary's average risk- and specialty-

adjusted monthly cost for a beneficiary during the measurement period is $1,187. The mean of beneficiary's average risk- and specialty-adjusted monthly cost for beneficiaries with services relating to acute inpatient admissions is $2,647, compared with $866 for a beneficiary without acute inpatient admissions. The mean of beneficiary's average risk- and specialty-adjusted monthly cost with services relating to Post-Acute Care is $2,427 compared with $996 for a beneficiary without PAC. This indicates that the measure can capture higher resource use by clinicians who have higher rates of complications related to these types of services.

The results from the clinical themes analysis found the correlation with risk- and specialty-adjusted cost were low to moderate. At both the TIN and TIN-NPI levels, there is a moderate correlation between the SNF service category and risk-adjusted cost (0.54), low correlation between Outpatient E&M Services, Procedures, and Therapy and risk-adjusted cost (0.45) and Acute Inpatient Services (0.38), and very low for the HH category (0.11), Non-Hospital Admission Emergency Services (0.15).

Finally, testing on attribution validity showed the measure is appropriately identifying and attributing beneficiaries to clinicians who have a primary care relationship with them. The mean share of beneficiary's E&M claims billed by attributed TINs or TIN-NPI is 52.8 percent and 45.0 percent, respectively, indicating attributed TIN/TIN-NPIs bill a somewhat significant proportion of beneficiaries' E&M claims related primary care as was intended, indicating a moderate relationship between attributed TIN/TIN-NPIs and beneficiaries they treat.

The analysis for attribution as a result of NP/PA show that logic attributes few TINs to providers who provide primary care largely through NPs and PAs. 13.3% of all TINs were attributed based on services conducted by NPs and/or PAs exclusively and 7.8% percent came from TINs comprised of a majority of NP and/or PAs. For TINs with specialties not primary consisting of NP or PA, only 5.5 percent of TINs were attributed via this method.

Panel Member #3: Face validity was good (4.8 out of 6); empirical validity (Pearson correlations) were very uneven—see following table.

Table 5. Pearson Correlation Statistics between Costs of Clinical Themes with Risk-Adjusted Cost

Clinical Theme	Pearson Correlation	
	TIN	TIN NPI
Acute Inpatient Services	0.38	0.38
Emergency Services Not Included in Hospital Admission	0.15	0.15
Outpatient E&M Services, Procedures, and Therapy	0.45	0.45
Post-Acute Care: Home Health	0.11	0.11
Post-Acute Care: IRF/LTCH	0.18	0.18
Post-Acute Care: SNF	0.54	0.54

Attribution was 50% or less (coin flip).

Table 6. Share of Primary Care E&M Billed by Attributed TIN and TIN-NPI

Provider Type	Distribution of Share of Attributed Beneficiary s Primary Care Related E&Ms Billed by Attributed TIN or TIN NPI			
	Mean	25th Percentile	Median	75th Percentile
TIN	52.8%	34.8%	55.8%	70.0%
TIN-NPI	45.0%	28.9%	42.8%	61.6%

Panel Member #5:

Panel Member #7: Distribution of beneficiary's average risk- and specialty-adjusted monthly cost were provided. Inpatient admissions and post-acute care were the major contributors to total cost. Pearson correlations for both TIN and TIN-NPI showed a positive correlation with TCC, highest for post-acute care SNF, outpatient E&M and inpatient care, all more than 0.38

Panel Member #9: Face validity survery results positive, 80% of respondents agree that measure scores would provide an accurate refletion of 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.

Share of E & M results are better for TIN than TIN-NPI. Attribution validity seems to be better for TIN than TIN-NPI.

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

    ☐ **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 <u>are</u> threats to validity and/or relevant threats to validity were <u>not assessed OR</u> 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 <u>is required</u>; 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 explanation for modelrate rating in 21 and 22 above.

    Panel Member #2: Face validity results showed moderate to strong validity related to differentiating beneficiary costs. Empirical validity testing . Empirical validity testings shws risk and speciality adjusted costs were significantly higher for beneficiaries with acute admissions and PAC as hypothesized. The clinical themes analysis show moderate to low correlations bewttwen costs and high cost service dcategories. The attribution logic validation shows patients are being attributed to TINs providing most of their primary care.

    Panel Member #3: Results for validity were underwhelming.

Panel Member #7: The model shows that total cost of care correlates highly with high costs of care in the subcomponents of care.

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.**

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 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.**
- **Uses existing claims data and codes.**
- **None.**
- **yes, this is feasible to implement. CMS is already doing so**
- **No concerns with generating the measure, but it will be challenging for clinicians to track and/or improve their performance more regularly.**
- **Think the availability of social risk data that would improve risk adjustment not yet routinely captured**
- **This measure is feasible to collect and calculate.**
- **No concerns**
- **Data elements (healthcare claims) are generated in the routine delivery of healthcare**

Criterion 4: Usability and Use

Use

4a. Use. evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

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 current TPCC measure, (ii) while the field test reports and other supplementary materials were helpful, the complexity of these documents was a challenge to some stakeholders, and; (iii) general feedback on the measure's attribution methodology, candidate events, and specialty adjustment.

Additional Feedback:

- This measure was reviewed by the Measure Applications Partnership (MAP) for 2018 – 2019 meausures under consideration.
- The MAP did not support this measure for rulemaking with the potential for mitigation. Mitigating factors include greater transparency around the attribution model and testing results.
- The MAP noted that this measure is an updated version of the total per capita cost measure currently used in MIPS and the potential updates include changes in the attribution methodology. The MAP raised concerns about the lack of available information on the measure's validity testing.
- The MAP also noted a need to better understand how this measure handles the issue of small numbers and evaluate if there is a need to include social risk factors in the measure's risk adjustment model.
- Finally, the MAP 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.

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *How has the measure been vetted in real-world settings by those being measured or others?*

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

Usability

4b. Usability.

The extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement.

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

4b2. Benefits vs. harms.

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).

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 for 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

- The measure, including the clinical and construction logic for a defined unit of measurement, can be deconstructed to facilitate transparency and understanding.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *What benefits, potential harms or unintended consequences should be considered?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Staff preliminary rating for Usability and Use: ☐ 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.**
- **Unclear.**
- **Yes. See the information provided by the NQF staff to this (4a1).**
- **yes, CMs will be using in CY 2022 as part of MIPS program--so providers will receive financial incentives based on their performance**
- **No comment.**
- **Developer claims publicly reported though unclear where as QPP reporting and Physician Compare are inconsistent in terms of being publicly accessible. Theoretically – as data files are made available. Unclear whether clinicians/groups can rework measures or use data to improve considering the complexity of the measure**
- **This measure is not publicly reported but is shared with MIPS participants.**
- **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**
- **Satisfied with the summary on the above questions compiled by the NQF staff (see section 4a2).**
- **yes, it was reviewed by the MAP which raised a number of concerns. It was also reviewed by experts (per face validity check)**
- **No concerns.**
- **Unclear. Report cites MAP review that did not support the measure due to lack of transparency and concerns with duplicative nature of having both TPCC and MSPB measures.**
- **n/a**
- **No concerns**

- **No concerns**

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.**
-
- **The developer has provided evidence that the implementation of this measure will reduce spending variability across Medicare physician or physician groups with regard to TPCC.**
- **will be used as part of MIPS. Usability is somewhat limited as this is summary measure and it doesn't provide drill down on cost drivers for physicians to quickly identify areas needing change. Onus is back on individual physician to sort this out (which takes time, money)**
- **Yes.**
- **No, which is concerning considering how complicated the measure attributions are – which make it unclear how clinicians are meant to understand the measure and the measure units in order to use the measure to improve.**
- **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.**
- **None.**
- **concern about not including social risk factor adjustments as this moves into determining payments under MIPS. This is a high stakes application. Need to adjust for within provider disparities due to social risk factors.**
- **Cost of care performance is frequently misinterpreted as representing "value." Any representation of performance on cost measures should be accompanied by information on quality performance.**

- **I think 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.**
- **The measure could be more beneficial is the developers explored appropriate resource use and its relationship to health outcomes.**
- **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):
    - 1604 : Total Cost of Care Population-based PMPM Index

Harmonization

- The measure developer indicates that this measure is not harmonized.
- The developer states that #1604 is tested and endorsed for a population of patients less than 65 years of age, while TPCC was developed and tested on the Medicare population, affecting the appropriate intended use of each respective measure.

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:

- **Discussed by developer.**
- **No competing measure identified. Related NQF-endorsed measure identified by the developer is: 1604 : Total Cost of Care Population-based PMPM Index . The measure developer indicates that this measure is not harmonized. The developer states that #1604 is tested and endorsed for a population of patients less than 65 years of age, while TPCC was developed and tested on the**
- **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.**
- **Interaction with the MSPB 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 TPCC measure captures the same costs as the episode-based measures, effectively "double counting" the costs, as is also true of the MSPB measure.**
- **none.**
- **No concerns**
- **Non**

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
  8409 :
  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 attribution methodology for this measure. While The Centers for Medicare and Medicaid Services (CMS) developed the Total Per Capita Cost (TPCC) 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 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 falls to do
- An unreliable attribution methodology that has several potential unintended consequences including, that multiple clinicians can be attributed to the measure unrelated to practicing team-based care.
- Neurologists and neurology advanced practice providers often consult with primary care and other specialists as they care for patients with neurological conditions, many of them chronic. Without clearer attribution methodology, the measure as written could give the perverse incentive to any of the clinicians involved to schedule follow up visits within three months so the patient and his or her costs are not attributed to that clinician.

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 Society of Retina Specialists (ASRS)
  Member Vote N/A
  8406 :
  The American Society of Retina Specialists appreciates the opportunity to provide comments on the Standing Committee related to its consideration of the Total Per Capital Cost (TPCC) measure. ASRS is the largest retinal organization in the world, representing over 3,000 members in all 50 US states, the District of Columbia, Puerto Rico, and 63 countries. The Society serves as a national advocate and primary source of clinical and scientific information and education for its members.

  ASRS opposes the inclusion of TPCC in the Merit-Based Incentive Payment System (MIPS) and urges the Standing Committee not to recommend it for endorsement because it potentially holds physicians responsible for the cost of care that they did not provide. While we continue to oppose the measure concept overall due to its attribution of costs, we appreciate that CMS has taken steps to target this measure more specifically to primary care physicians by excluding most specialists and surgeons, including all ophthalmologists. If this measure is not removed from the MIPS program, however, at a very minimum it must retain the specialty exclusions.

  This measure is not appropriate for measuring or influencing individual physician performance and thus should not be included in the MIPS program. TPCC seeks to assign the total costs of all care for individual Medicare beneficiaries to primary care physicians tasked with coordinating and overseeing the total healthcare of the patient. While CMS listened to stakeholder feedback that the previous attribution methodology for the measure potentially held physicians responsible for care that happened before the primary care physician saw the patient or well after the patient had left that particular physician's care, its revisions to the methodology to shorten the time frame for attribution have created an additional problem of potentially double-counting costs by assigning them to more than one physician or group. This risks confusion over who is responsible for the costs and may inappropriately label a physician as high or low cost.

  Furthermore, the measure updates fail to address the underlying issue that the attributed physician is still being held responsible for the cost of care that he or she neither provides nor has any ability to influence. Under the TPCC methodology, a primary care physician will be attributed the cost of

care provided by a retina specialist, such as macular degeneration treatment or surgical repair of a torn or detached retina, even though he or she has no influence over the cost or quality of that treatment. By including all costs of care for a particular beneficiary, the measure loses its overall usefulness to the attributed physician since he or she is limited in the ability to modify or influence the behavior of other physicians caring for the patient. The physician will not ultimately take action that lowers the cost of care for the beneficiary and the measure score will not accurately assign physicians as high or low cost.

Despite ASRS' overall opposition to the measure concept, we applaud CMS for listening to feedback and excluding specialists and surgeons, including all ophthalmologists, who had patients inappropriately attributed to them by application of the TPCC measure under its prior methodology. Previously, beneficiaries who did not see a primary care physician sometime during the performance year were attributed to whichever physician or group billed the plurality of evaluation and management (E/M) services during the year, which could be a retina specialist. Retina specialists provide care only for diseases of the retina and macula and do not provide overall or systemic healthcare for the patient. While retina specialists treat diabetic eye disease, such as diabetic retinopathy, they do not manage the patient's overall diabetes care. Although any physician is limited in his or her ability to influence their TPCC score, retina specialists are especially disadvantaged since the care they provide is so specialized. The new attribution methodology appropriately excludes them and all other ophthalmologists from the measure and must be retained if this flawed measure is included in the MIPS program.

Thank you for the opportunity to provide comments on the TPCC measure. ASRS continues to oppose the inclusion of this measure in the MIPS program because it holds physicians responsible for the cost of care they did not provide, thereby limiting their ability to influence their performance on the measure. While this measure should be removed from MIPS entirely, retention of the specialty exclusions is necessary to ensure specialists and surgeons, such as retina specialists, do not have patients and costs inappropriately attributed to them under the measure. We urge the Standing Committee not to recommend this measure for endorsement.

For additional information, please contact Allison Madson, director of health policy, at allison.madson@asrs.org or (312) 578-8760.

- Comment by Infectious Diseases Society of America (IDSA)
  Member Vote N/A
  8405 :
  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
  8403 :
  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.

  ASCO requests that the Standing Committee evaluate whether the attribution methodology is valid and does not lead to negative unintended consequences. While the TPCC eliminates the problem of attributing costs that occurred before the clinician ever saw the patient, which ASCO supports ,the attribution methodology assumes that a primary care relationship exists if two things happen within three days or three months, and not otherwise, leading to problems as identified in the following examples:

  A cancer survivor receives a twelve-month follow-up exam from their oncologist, along with a two-dimensional echocardiogram with doppler flow study to screen for cardiotoxicity. The oncologist is attributed the beneficiary's costs for a twelve-month period, despite no other management of the patient.

  A newly diagnosed cancer patient requests a second opinion from an oncologist other than their primary clinician. The oncologist conducts an evaluation and management service which happens to take place within +/- 3 days of other designated primary care services. The oncologist performing the second opinion confirms the primary clinician's treatment plan and the patient returns to their primary clinician for continued management and treatment. The consulting oncologist is attributed the beneficiary's costs for a twelve-month period, despite never having managed the beneficiary's care.

  A nurse practitioner is employed by a cancer practice to assist in management of cancer patients receiving chemotherapy and/or radiation therapy. The nurse practitioner does not qualify for an exemption from the measure given that a physician's NPI, rather than theirs, is used to bill for the chemotherapy services.

An oncologist whose TIN includes in-office chemotherapy services is attributed a patient who receives chemotherapy services within 90 days after an E&M primary care service, but outside of +/- 3 days of the E&M primary care service. An equivalent hospital-based oncologist is not attributed a similar case, as the chemotherapy services are billed by the hospital TIN.

An oncologist whose TIN includes in-office chemotherapy services qualifies for an exemption from the measure due to their NPI-TIN being used to bill for chemotherapy services. An equivalent hospital-based oncologist does not qualify for an exemption, as the chemotherapy services are billed by the hospital TIN.

In each of these examples, an oncologist will not know if they qualify for the TPCC measure, as the exemption is applied retrospectively based on a measurement of candidate events for which the oncologist bills for chemotherapy or radiation therapy services. We feel it is inappropriate for a clinician to be included in a measure for which they are unaware of which beneficiaries they may be attributed, or whether they will receive an exemption. We have previously recommended that all medical and radiation oncologists be excluded from the TPCC measure.

The analysis of the extent to which nurse practitioners and physician assistants have this measure attributed to them found that 5.5% of practices (just over 4,000 of the 74,191 TINs) were ones in which the specialty is not considered to provide primary care. We believe that these findings are the result of the decision to make exclusions at the specialty level and not at the service level. While the measure excludes certain specialties, the results as outlined in Table 7 (Frequency of Most Common HCFA Specialties in TINs Attributed TPCC Measure via Nurse Practitioners and Physician Assistants Alone) of the testing form confirm that there is potential for the measure to attribute patients to clinicians who do not provide primary care services. These results from both analyses lead to questions on the validity of the attribution methodology as it creates a fairness issue by sometimes including certain specialties regarded as not providing primary care, but it also holds primary care clinicians responsible for the costs of non-primary-care services that they do not provide and cannot control.

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

- Comment by American College of Physicians (ACP)
  Member Vote N/A
  8398 :
  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 Total per Capita Cost measure represents an important move towards cost assessment in pay-for-performance programs. However, the methods that policymakers and measure developers apply to assessing 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 beneficiary to a single Taxpayer Identification Number-National Provider Identifier (TIN-NPI) if the beneficiary received more primary care services from primary care clinicians in that TIN-NPI than any other TIN-NPI or CMS Certification Number (CCN). If two TIN-NPIs tie for the largest share of a beneficiary's primary care services, CMS attributes the beneficiary to the TIN-NPI that provided primary care services most recently. According to this model, multiple clinicians could be accountable for the annualized costs of care for beneficiaries attributed to the TIN-NPI. While it is reasonable to apply this model to health plans, it is unclear how this approach will provide meaningful information to individual clinicians that will appropriately inform quality improvements. While we generally support the attribution model at the facility, system, and health plan levels, we caution CMS 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:

- 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.
    - 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.
    - Additionally, 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.
    - CMS should independently establish a robust minimum average reliability rating and evaluate all future cost measures based on that same standard, not pre-determine a set of measures the Agency wishes to use then selecting whatever low reliability standard allows them to adopt all of those measures without raising case minimums.
    - CMS designed this measure to seemingly reward the creation of Patient-Centered Medical Homes; however, PCMH models have not been uniformly successful in achieving care quality improvements.

- Comment by Federation of American Hospitals (FAH)
  Member Vote N/A
  8393 :
  The Federation of American Hospitals (FAH) appreciates the opportunity to comment on this measure prior to the Standing Committee's evaluation. The FAH requests that the committee carefully consider whether the measure as specified produces performance scores that are reliable and valid for reporting at either the clinician or practice levels.

  While reliability at the 25th percentile for at least 20 episodes for practices was 0.77 and 0.83 for clinicians, the FAH questions what result was produced at the minimum level for either reporting group. The FAH was particularly concerned to review the additional explanation provided in section 2a2.3 in the testing form that "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", which should not be considered an acceptable minimum threshold. The FAH believes that additional information regarding the minimum result is needed to ensure that the measure as specified produces scores that achieve an acceptable minimum threshold for reliability.

  The FAH is extremely troubled by the lack of any validity testing demonstrating the presence or absence of correlations of this cost measure to quality measures. We found the rationales outlining the inability of the developer to identify appropriate quality measures to be weak since QPP#458, All-cause Hospital Readmission, which is also a claims-based measure, was attributed to groups in 2017 and we assume that CMS would be able to enable matching of groups to whom this measure and QPP#458 applied. In addition, it is concerning for a cost measure to be considered for endorsement when the developer is unable to identify applicable quality measures and reports that the quality measures used within the program do not enable this type of analysis; yet, points are achieved and an overall score is derived in part from the quality and cost categories in MIPS. Due to the importance of understanding costs as they relate to quality, it seems imprudent to consider endorsement of a measure for which its association to quality cannot be demonstrated.

The FAH was further concerned with the results from the analysis of the measure's attribution methodology. Similar to our questions on the reliability testing, the 25th percentile showing that just under 29% and 35% of E&M claims were billed by the attributed clinician and practice, respectively. These results lead us to ask what share of claims occurred at lower levels. The other analysis to determine the extent to which nurse practitioners and physician assistants to whom this measure was attributed also shows that inappropriate attributions are occurring to 5.5% of practices. While the developer views this result to be acceptable, the specialties in which this occurs includes those that are intentionally excluded within the specifications. We believe that these results impact the validity of the measure and could result in negative unintended consequences in attributing costs to clinicians and practices who do not provide primary care services.

In addition, while the FAH appreciates that social risk factors were reviewed, and believes that the risk adjustment approach should not consider the identification and testing of social risk factors as supplementary to clinical risk factors. This approach was identified as a concern by the NQF Disparities Standing Committee and developers must begin to include these factors within the testing of the model rather than the approach of "adding on" factors after the model is developed. This type of analysis would assist facilities and others in understanding how their inclusion could impact the model and provide additional information for groups examining this issue such as the NQF and Office of the Assistant Secretary for Planning and Evaluation. As a result, the FAH believes that this measure lacks sufficient information on the potential impact these social risk variables have on the risk adjustment model.

Furthermore, while the developer provides information on the differences in observed to expected cost ratios that result from social risk factor adjustment (table 2b3.4b.c), it is not necessarily clear on the degree to which these changes would result in a practice or clinician's result changing. Specifically, what is not answered is whether the addition of social risk factors to the model would lead to a clinician or practice earning higher or lower points in the benchmarks currently used for this measure within MIPS. If the interpretation of the results under meaningful differences leads to statements that "small differences in scores can be interpreted as meaningful"(response to 2b4.3 in the testing form), to what extent would changes in performance as a result of adjustment for social risk factors also lead to different but meaningful results?

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)

Effective primary care management can support Medicare savings in several ways. For example, more effective primary care management can improve the treatment of chronic conditions by obviating the need for high-cost hospital or emergency department services. It can also direct a greater proportion of patients to lower hospital costs for inpatient services. [1] Given the potential for decreasing spending through improvements in primary care delivery, the TPCC measure allows for a savings opportunity by capturing the broader healthcare costs influenced by primary care.

[1] "Valuation of Care Management Performed by Primary Care Services: An Issue Brief." American Academy of Family Physicians, 2018.

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 74,191 clinician group practices (identified by Tax Identification Number [TIN]) and 335,480 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 20 or more TPCC beneficiaries, as identified in Medicare Parts A and B claims data, during January 1, 2018, to December 31, 2018. Beneficiaries from all 50 States and D.C. receiving evaluation and management care indicative of primary care were included, with their respective costs evaluated from all claim settings.

TIN Level Scores:

- Mean score: $1,109
- Standard deviation: $257
- Min score: $35
- Max score: $8,449
- Score IQR: $255
- Score percentiles
  - 10th: $833
  - 20th: $935
  - 30th: $999

- 40th: $1,049
- 50th: $1,095
- 60th: $1,141
- 70th: $1,192
- 80th: $1,262
- 90th: $1,383

- Number of beneficiaries: 26,636,602

TIN-NPI Level Scores:

- Mean score: $1,169
- Standard deviation: $310
- Min score: $7
- Max score: $10,024
- Score IQR: $297
- Score percentiles
  - 10th: $855
  - 20th: $961
  - 30th: $1,030
  - 40th: $1,087
  - 50th: $1,140
  - 60th: $1,194
  - 70th: $1,257
  - 80th: $1,341
  - 90th: $1,489
- Number of beneficiaries: 26,374,993

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.

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

Rationale for Measuring Cost through All-Cost Measure vs. Episode-Based Cost Measure

TPCC is a broad measure that focuses on measuring the performance of clinicians delivering primary care services, which can include both primary care and specialty clinicians. By allowing more clinicians to have their cost performance measured, this broad measure complements more specific episode-based cost measures, which measure the performance of a subset of specialties concentrated around a specific condition or procedure. In complementing episode-based cost measures, all-cost measures, such as TPCC, become an important means to enhance the coverage of patients and effectively incentivize improvements in the efficiency of care delivery in Medicare. Inclusions of all costs provides a broad assessment of a clinician's management of the overall health of a patient, as opposed to episode-based cost measures, which only capture clinically-related services for a given procedure or condition. In managing a patient's complete health, clinicians measured under the TPCC measure are incentivized to conduct patient follow-up, coordinate care amongst specialists, offer necessary referrals, and actively diagnose patients.

Rationale for Measuring the Total Per Capita Cost (TPCC)

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.

Research shows that primary care management in certain settings, such as Patient-Centered Medical Homes (PCMH), has brought about measurable reductions to the total cost of care by reducing utilization of high-cost services and in some cases, by directing patients to lower cost hospitals. [3] With this research-based evidence available for certain settings, a key question for policymakers is whether primary care management would achieve similar results across a wider variety of settings. In light of this question, a measure that captures the cost performance of primary care providers across a range of settings can help to confirm the benefits of effective primary care management. Given that, as noted above, clinicians are often unaware of how their choices affect the total costs of care, such a measure can help guide primary care providers towards practices that reduce costs, while maintaining or improving quality.

Another key opportunity presented by a cost performance measure for primary care is the opportunity to reward primary care providers for delivering value and to thereby improve patients' access to primary care services. As noted by MedPAC, beneficiaries experience more difficulty accessing primary care than with accessing specialty care. [4] More specifically, 1.3 percent of the Medicare population reported a "big problem" finding a primary care doctor, while just 0.9 percent of this population reported such a problem in finding a

specialist in 2017. Relatedly, among patients desiring to switch primary care providers, some patients felt that this was not an option due to long wait times or due to practices being closed to new patients. This may be related to another fact that MedPAC observes in the same report, which is that the Physician Fee Schedule's orientation to discrete services with a clear beginning and end does not support primary care, with its need for ongoing care coordination for a group of patients. Given this, MedPAC recommended the establishment of a per beneficiary payment for primary care practitioners to replace the expired Primary Care Incentive Payment (PCIP) program. This program provided a 10 percent bonus on fee schedule payments for some E&M services delivered by primary care practitioners. While the establishment of such a revised payment policy for primary care management might be an optimal solution to increase the availability of primary care, it may take substantial time to implement. Given this, it is particularly important to utilize an existing measure of the cost performance of primary care clinicians to identify and provide financial incentives for good performance.

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 TPCC 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 TPCC cost performance.

[1] CMS, "Merit-Based Incentive Payment System (MIPS): Total Per Capita Cost (TPCC) Measure – Measure Information Form," https://www.cms.gov/files/zip/2020-cost-measure-information-forms.zip.

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

[3] "Valuation of Care Management Performed by Primary Care Services: An Issue Brief." American Academy of Family Physicians, 2018.

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

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):*

No Applicable Care Setting

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 | 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 capita (population- or patient-based)

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: TPCC uses Part A and B claims data to attribute beneficiaries to clinicians, calculate beneficiary's 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; other primary payers; disability status; sex; end-stage renal disease (ESRD); beneficiary birth dates; and beneficiary death dates.

Common Medicare Environment (CME) database: This is used to determine beneficiary's dual status. https://www.ccwdata.org/documents/10280/19002256/medicare-enrollment-impact-of-conversion-from-edb-to-cme.pdf.

Minimum Data Set (MDS): The MDS is used to identify beneficiaries that should be risk adjusted through the CMS-HCC v22 institutional model.

https://www.resdac.org/cms-data/files/mds-3.0.

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

https://www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html.

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_tpcc.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-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 TPCC measure score is calculated as the average payment-standardized, risk-adjusted, and specialty-adjusted monthly costs across all beneficiary months in the performance period attributed to a clinician group (TIN) or individual clinician (TIN-NPI).

The measure population is identified as beneficiaries for whom a clinician group (TIN) provides two outpatient 'primary care services' within 90 days and to the individual clinician (TIN-NPI) within a TIN that provides the most 'primary care evaluation and management services'. Certain types of clinicians are excluded from attribution if their practice patterns identified through claims billing focus on global surgery, anesthesia, therapeutic radiation, or chemotherapy. Certain specialties that are not reasonably responsible for providing primary care are also excluded based on their HCFA Specialty designation. Beneficiaries are attributed for a one year period and measured on a monthly basis for those months occurring during the performance period. The costs of all Part A and Part B services occurring during the attributed beneficiary's months are summed to obtain each month's standardized observed cost. The monthly observed costs are risk-adjusted by dividing by the beneficiary's risk scores as determined by the CMS-HCC and CMS-ESRD risk adjustment models for patient characteristics found in the year prior to that particular month. A specialty adjustment is then applied to monthly risk-adjusted costs to account for the fact that costs vary across specialties and across TINs with different specialty compositions.

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: Identify Beneficiaries for Attribution (i.e., Candidate Events)

A 'candidate event' indicates the start of a primary care relationship between a clinician and beneficiary, and is identified by the occurrence of two Part B Physician/Supplier (Carrier) claims with particular CPT/HCPCS services billed in close proximity. There are two different sets of CPT/HCPCS codes used: E&M primary care services and primary care services.

E&M primary care services are a specific set of evaluation and management codes for clinician visits in the outpatient setting, physician office, nursing facility, or assisted living.

Primary care services are a broader list of services related to routine primary care and generally fall into the following categories: Durable Medical Equipment (DME) and Supplies, Electrocardiogram, Laboratory - Chemistry and Hematology, Other Diagnostic Procedures (Interview, Evaluation, and Consultation), Other Diagnostic Radiology and Related Techniques, Prophylactic Vaccinations and Inoculations, Routine Chest X-ray, Clinical Labs, Preventive Services.

To identify a candidate event, firstly, an initial E&M primary care service billed on Part B Physician/Supplier (Carrier) claim is identified. This E&M primary care service is not considered if it occurs during a beneficiary's

stay at a Critical Access Hospital (CAH), Inpatient Facility, or Skilled Nursing Facility (SNF). Secondly, in addition to the initial E&M primary care service, the presence of at least one of the following services confirms the candidate event:

•From any TIN within +/- 3 days: Another primary care service,

•From the same TIN within + 90 days: A second E&M primary care service OR another primary care service

See the "Prim_Care_E&Ms" and the "Prim_Care_Services" tabs of the TPCC Measure Codes List file for the list of the Current Procedural Terminology/Healthcare Common Procedure Coding System (CPT/HCPCS) codes that identify E&M primary care services and primary care services, respectively. The URL of this downloadable file is linked in Section S.1.

STEP 2: Exclude Clinicians Unlikely to be Providing Primary Care

Once candidate events are identified, individual clinicians (identified by TIN-NPI) can be attributed based on their involvement in the candidate event and how their practice relates to primary care. The TIN-NPI assigned responsible for a candidate event is the clinician found on the initial E&M primary care service claim of the candidate event. TIN-NPIs are excluded from attribution if they meet one of two types of exclusions: service category exclusions and specialty exclusions. Candidate events belonging to TIN-NPIs who meet any of these exclusions are removed from attribution and measure calculation for both the TIN-NPI and their respective clinician group (identified by TIN).

STEP 2.1: Exclude Clinicians Based on Service Category Exclusions

Clinicians whose billing patterns indicate that they tend to provide services that are not within the scope of primary care are excluded from attribution of the TPCC measure. A TIN-NPI and all their candidate events are removed from attribution if he or she bills the volume of services below within +/-180 days of a candidate event for a beneficiary:

•At least 15 percent of the clinician's candidate events are billed with 10-day or 90-day global surgery services.

•At least 5 percent of the clinician's candidate events are billed with anesthesia services.

•At least 5 percent of the clinician's candidate events are billed with therapeutic radiation services.

•At least 10 percent of the clinician's candidate events are billed with chemotherapy services.

The list of CPT/HCPCS codes used for each of the service exclusions can be found in the tabs of the TPCC Measure Codes List file labeled: "HCPCS_Surgery," "HCPCS_Anesthesia," "HCPCS_Ther_Rad," and "HCPCS_Chemo." The downloadable file is linked in Section S.1.

STEP 2.2: Exclude Clinicians Based on Specialty Exclusions

Clinicians who – based on their specialty – would not reasonably be responsible for providing primary care are excluded from attribution of the TPCC measure. This exclusion aims to keep primary care specialists and internal medicine subspecialists who frequently manage patients with chronic conditions falling in their areas of specialty. The excluded specialties list contains 56 specialties that fall into the following broad categories:

•Surgical sub-specialties

•Non-physicians without chronic management of significant medical conditions

•Internal medicine sub-specialties with additional highly procedural sub-specialization

•Internal medicine specialties that practice primarily inpatient care without chronic care management

•Pediatricians who do not typically practice adult medicine

The list of HCFA Specialty codes that identify clinicians that are included or excluded from the measure attribution can be found in the "Eligible_Clinicians" tab of the TPCC Measure Codes List. The downloadable file is linked in Section S.1.

STEP 3: Construct Risk Windows

Candidate events that are not excluded initiate the opening of a risk window, a year-long period that begins on the date of the initial E&M primary care service of the candidate event. The performance period is divided into 13 four-week blocks called beneficiary months. Beneficiary months during the risk window are considered attributable if they occur during the performance period. In the event that a risk window begins or ends with a partially covered month, only the portion during the risk window and the performance period is considered for attribution.

STEP 4: Attribute Beneficiary Months to TINs and TIN-NPIs

Beneficiary months are attributed to a TIN according to the following steps:

•Identify the TIN billing the initial E&M primary care service claim of each candidate event.

•Determine beneficiary months that fall within the risk windows of the candidate events that were initiated by the TIN and overlap the performance period and attribute those beneficiary months to the TIN.

Beneficiary months are attributed to a TIN-NPI according to the following steps:

•Identify the TIN-NPI billing the initial E&M primary care service claim of each candidate event.

•Determine beneficiary months that fall within the risk windows of the candidate events that were initiated by the TIN-NPI and that overlap the performance period.

•Identify the TIN-NPI within an attributed TIN that is responsible for the plurality of candidate events provided to the beneficiary. If two or more TIN-NPIs under a single TIN provide the same proportion of candidate events to a beneficiary, attribute the beneficiary to the TIN-NPI that provided the earliest candidate event.

•Attribute only the beneficiary months from candidate events that the TIN-NPI is responsible for initiating, which is not necessarily all candidate events attributed to the TIN for that beneficiary.

STEP 5: Calculate Payment-Standardized Monthly Observed Costs

The monthly observed cost for attributed beneficiary months is the sum of all service costs billed for a particular beneficiary during a beneficiary month. Monthly observed costs are standardized to account for differences in Medicare payments for the same service(s) across Medicare providers. Payment standardization accounts for differences in Medicare payment unrelated to the care provided, such as those from payment adjustments supporting larger Medicare program goals (e.g. indirect medical education add-on payments) or variation in regional healthcare expenses as measured by hospital wage indexes and geographic price cost indexes (GPCIs). Standardized costs that occur during partially covered months are pro-rated, based on the portion of the month covered by the risk window.

STEP 6: Risk-Adjust Monthly Costs

Beneficiary cost may differ across clinicians for reasons unrelated to the attributed clinicians' treatment and outside of their control. Risk adjustment accounts for case-mix of patients and other non-clinical characteristics that influence complexity of case-mix and is defined by a patient's claims found one year prior the start of a respective beneficiary month. The CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) 2016 Risk Adjustment models are used for beneficiaries without End Stage Renal Disease (ESRD). Specifically,

•The new enrollee model is used for beneficiaries that have fewer than 12 months of Medicare medical history. The model accounts for each beneficiary's age, sex, disability status, original reason for Medicare entitlement (age or disability), and Medicaid eligibility.

•The community model is used for beneficiaries that have least 12 months of Medicare medical history. The model includes the same demographic information as the new enrollee model but also accounts for clinical conditions as measured by HCCs.

•The institutional model is used for beneficiaries who were in long-term institutional settings. The model includes demographic variables, clinical conditions as measured by HCCs, and various interaction terms.

The CMS-ESRD Version 21 (CMS-ESRD V21) 2016 Risk Adjustment models are used for ESRD beneficiaries receiving dialysis. Specifically,

•The dialysis new enrollee model is used for ESRD beneficiaries that have fewer than 12 months of Medicare medical history. The model accounts for each beneficiary's age, sex, disability status, original reason for Medicare entitlement (age or disability), Medicaid eligibility, and ESRD.

•The dialysis community model is used for ESRD beneficiaries that have at least 12 months of Medicare medical history. The model includes the same demographic information as the new enrollee model but also accounts for clinical conditions as measured by HCCs.

The "HCC_Risk_Adjust" tab of the Measure Codes List file lists all variables included in the CMS-ESRD V21 and the CMS-HCC V22 risk adjustment models. The downloadable file is linked in Section S.1.

The standardized risk scores from the CMS-ESRD V21 and CMS-HCC V22 models are generated for each beneficiary's month that summarizes the beneficiary's expected cost of care relative to other beneficiaries. Risk scores for ESRD beneficiaries are normalized to be on a comparable scale with the HCC V22 risk scores. A risk score equal to 1 indicates risk associated with expenditures for the average beneficiary nationwide. A risk score greater than 1 indicates above average risk, while a risk score less than 1 indicates below average risk.

The risk-adjusted monthly cost for each attributed month is calculated according to the following steps:

•Calculate CMS risk score for each beneficiary month using diagnostic data from the year prior to the month. This risk score is normalized by dividing by the average risk score for all beneficiary months.

•Divide observed costs for each beneficiary month by the normalized risk score to obtain risk-adjusted monthly costs.

•Winsorize risk-adjusted monthly costs at the 99th percentile by assigning the 99th percentile of monthly costs to all attributable beneficiary months with costs above the 99th percentile.

•Normalize monthly costs to account for differences in expected costs based on the number of clinician groups to which a beneficiary is attributed in a given month. The normalization factor is the inverse cube root of the number of attributed clinician groups for that beneficiary month.

STEP 7: Specialty-Adjust Monthly Cost

The specialty adjustment for the TPCC measure is a cost adjustment applied to account for the fact that costs vary across specialties and across TINs with varying specialty compositions. The specialty adjustment at the TIN and TIN-NPI levels is calculated as follows:

1) Calculate the average risk-adjusted monthly cost for each TIN and TIN-NPI by averaging risk-adjusted monthly cost across all attributed beneficiary months.

2) Calculate the national specialty-specific expected cost for each specialty as the weighted average of TIN/TIN-NPI's risk-adjusted monthly cost.

2a) Define the weight for each TIN/TIN-NPI as the percentage of clinicians with that specialty multiplied by the total number of beneficiary months attributed to the TIN/TIN-NPI multiplied by the number of clinicians with that specialty.

2b) There will only be one specialty designation for a TIN-NPI. Therefore, the percentage of clinicians with a specialty and number of clinicians with a specialty will always be equal to 1.

3) Calculate the specialty-adjustment factor for each TIN or TIN-NPI as follows:

3a) Multiply the national specialty-specific expected cost for each specialty by the respective specialty's share of Part B payment within a TIN or TIN-NPI.

3b) Sum the weighted share of national specialty-specific expected cost calculated in the previous step across all the specialties under a given TIN or TIN-NPI.

STEP 8: Calculate the TPCC Measure Score

Calculate final risk-adjusted, specialty-adjusted cost measure by dividing each TIN and TIN-NPI's average risk-adjusted monthly cost by their specialty-adjustment factor and multiply this ratio by the average non-risk-adjusted, winsorized observed cost across the total population of attributed beneficiary months.

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

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 TPCC measure can identify the same beneficiary months to be attributed to the same clinician or clinician group as a result of separate overlapping candidate events (i.e., trigger events described in Section S.7.2. in Step 1). When this occurs, the measure will only include the beneficiary month once in the calculation for the respective clinician or clinician group.

The measure can attribute a beneficiary to multiple clinicians or clinician groups if evidence is found that both groups are managing the beneficiary concurrently. The measure calculation risk adjusts each clinician's observed costs for the patient with the same observable characteristics among their peers, rather than to a pre-defined standard. By comparing clinicians to their peers, who are all attributed in the same way, and measuring all clinicians who are responsible for the patient's care, we can expect this comparison to be fair. Allowing multiple clinicians to be attributed a beneficiary is an important feature of the measure as it ensures that all clinicians involved in a beneficiary's care are appropriately measured and subject to similar incentives, promoting joint accountability.

The measure accounts for disease interactions through its risk adjustment models specified by CMS' Hierarchical Condition Category Version 22 (CMS-HCC V22) and CMS' ESRD Version 21 (CMS-ESRD V21). In addition to the HCCs, the model includes disease interactions (e.g., Cancer * Immune Disorders). Further details about the risk adjustment models and disease interaction terms are included in Section S.8.6. and Section S.9.2.

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

Identification of a primary care relationship between a clinician and beneficiary are identified by the occurrence of two Part B Physician/Supplier (Carrier) claims with particular CPT/HCPCS services indicative of primary care. Specifically, evaluation and management codes for clinician visits in the outpatient setting, physician office, nursing facility, or assisted living, and a broader list of services related to routine primary care are used.

The TPCC measure includes all services during periods of attribution in the measurement. The TPCC captures a broad view of provider care, and focuses on measuring the performance of clinicians delivering primary care services. By not excluding services, this intends to capture overall costs of care, reflect the general management of beneficiary health that clinicians who provide primary care undertake, and incentivize

accountability for primary care clinicians to help protect against the diverse set of consequences for inappropriately managed diseases, missed diagnoses, or inappropriate specialty referrals.

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

The TPCC measure uses clinical indicators from CMS' Hierarchical Condition Category (HCC) and ESRD models, as described in Section S.7.2. Using different models for different types of patients helps to capture comorbidities that reflect particular patient profiles, such as patients in long-term care settings. This approach is adopted to ensure sufficient capture of the patient's comorbid disposition prior to the beneficiary month's cost being accessed to allow more comprehensive risk adjustment of comorbid factors. The Hierarchical Condition Categories prevent collinearity by suppressing HCCs for less severe manifestations of a conditions when evidence for the more severe condition is found. The general risk adjustment approach is detailed in Sections S.7.2. and S.9.3.

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 TPCC 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 included, the measure excludes beneficiary when the date of birth information (an input to the risk adjustment model) cannot be found in the EDB.

The TPCC measure also excludes beneficiary enrolled in Medicare Part C or has a primary payer other than Medicare at any point during the performance period. 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.

[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

All Part A and B

All Part A and B

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 TPCC measure focuses on primary care by design and includes all costs to provide a broad assessment of a clinician's management of the overall health of a patient, rather than a specific condition. In managing a patient's complete health, clinicians measured under the TPCC measure are incentivized to conduct patient follow-up, coordinate care amongst specialists, offer necessary referrals, and actively diagnose patients.

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 capture the overall costs of care to provide information to clinicians providing primary care services with the goal of incentivizing the provision of high-quality, cost-effective care. The clinical logic is constructed to achieve this objective.

Clinical Topic Area: Population-based measure for beneficiaries receiving primary care

Comorbidity and interactions: The risk adjustment models include a series of interaction terms between comorbidities. The risk adjustment models are also used to account for clinical severity levels of beneficiaries.

Clinical hierarchies: Clinical hierarchies are embedded within the risk adjustment models, and in determining which model applies to a given beneficiary.

Additional clinical logic includes accounting for the attribution of beneficiaries to clinicians through an evaluation of Part B services indicating primary care practice relationships, and ensuring that the measure is appropriately capturing clinicians who provide primary care services by excluding a defined set of clinicians either through their CMS HCFA specialty or Part B billing patterns.

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.)*

As described in Section S.7.2, to account for the clinical severity of patients, one of five separate risk adjustment models are applied based on the patients characteristics observed in the year prior to the beneficiary month being measured. For non-ESRD patients, the three models are the new enrollee model, community model, and institutional model from CMS' Hierarchical Condition Category Version 22 (CMS-HCC V22). For ESRD patients, the two models are the dialysis new enrollee model and dialysis community model from CMS' ESRD Version 21 (CMS-ESRD V21). Each model includes beneficiary demographic and enrollment information such as age, gender, disability, and dual enrollment status. Both the new enrollee model and dialysis new enrollee models are limited to these factors as the patient does not have sufficient Medicare claims history for further evaluation. The remaining models (community model, institutional model, and dialysis community) include either 79 (CMS-HCC V22) or 87 (CMS-ESRD V21) hierarchical condition categories to characterize the patient severity and comorbidities. The indicators used for risk adjustment and the methodology are detailed in the Measure Information Form linked in Section S.1.

The start of a primary care relationship between a clinician and beneficiary is identified by the occurrence of two Part B Physician/Supplier (Carrier) claims with particular CPT/HCPCS services billed in close proximity. There are two different sets of CPT/HCPCS codes used: E&M primary care services and primary care services. E&M primary care services are a specific set of evaluation and management codes for physician visits in the outpatient setting, physician office, nursing facility, or assisted living. Primary care services are a broader list of services related to routine primary care that generally fall into the following categories: Durable Medical Equipment (DME) and Supplies, Electrocardiogram, Laboratory - Chemistry and Hematology, Other Diagnostic Procedures (Interview, Evaluation, Consultation), Other Diagnostic Radiology and Related Techniques, Prophylactic Vaccinations and Inoculations, Routine Chest X-ray, Clinical Labs, and Preventive Services

The codes used to attribute beneficiaries to clinicians are listed in the tabs titled E&M_Prim_Care and Prim_Care_Services within the Measure Codes List linked in Section S.1.

Clinicians who would not reasonably be responsible for providing primary care are excluded from attribution of the revised TPCC measure using their CMS HCFA specialty designation assigned on Part B physician/supplier claims. This exclusion aims to keep primary care specialists and internal medicine subspecialists who frequently manage patients with chronic conditions falling in their areas of specialty. The excluded specialties list contains 56 specialties that fall into the following broad categories:

- Surgical sub-specialties
- Non-physicians without chronic management of significant medical conditions
- Internal medicine sub-specialties with additional highly procedural sub-specialization
- Internal medicine specialties that practice primarily inpatient care without chronic care management
- Pediatricians who do not typically practice adult medicine

The codes used to exclude clinicians from attribution base on their CMS HCFA specialty are listed in in the tab titled Eligible_Clinicians within the Measure Codes List linked in Section S.1.

Additionally, TIN-NPI are removed from attribution if a clinician met any of the following four service category thresholds for the same beneficiary by billing the specified CPT/HCPCS within +/-180 days of the candidate event on Part B physician/supplier claims:

•At least 15 percent of the clinician's attributable events are comprised of 10-day or 90-day global surgery services.

•At least 5 percent of the clinician's attributable events are comprised of anesthesia services.

•At least 5 percent of the clinician's attributable events are comprised of therapeutic radiation services.

•At least 10 percent of the clinician's attributable events are comprised of chemotherapy services.

The codes used to exclude clinicians from attribution base on Part billing patterns are listed in in the tabs titled HCPCS_Surgery, HCPCS_Anesthesia, HCPCS_Ther_Rad, HCPCS_Chemo within 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 TPCC 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 overall resource use for patients with a focus on clinician(s) providing primary care services. 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. One of the goals for using cost measures is to help inform clinicians on the costs attributable to their decision-making, as well as the total cost of their patient's care. A cost measure offers opportunity for improvement if clinicians can exercise influence on a significant share of costs during periods in which they can be considered responsible for a beneficiary, or if lower spending and better care quality can be delivered through changes in clinical practice. [1]

Physician services are an area of high spending where increased cost effectiveness can be impactful in reigning in Medicare spending: in 2017, Medicare FFS paid $69.1 billion for physician and other health professional services, accounting for around 14 percent of FFS Medicare spending.[4] Payment models like MIPS can have significant impacts on reducing costs and making care more affordable. Clinicians providing primary care can reduce the total cost of care by reducing utilization of high-cost services and in some cases, by directing patients to lower cost hospitals.[3] Small practice changes by all clinicians can have a sizable impact on reducing unnecessary healthcare spending; these findings led to the Choosing Wisely campaign that contains over 550 recommendations for unnecessary tests and treatments, with the participation of over 80 specialty societies.[8,9] Primary care clinicians have a role in minimizing the use of low value services where there is little or no clinical benefit or care where the risk of harm from the service outweighs the potential benefit [6,7,8]. Low-value services include unnecessary tests and treatment (e.g., imaging for non-specific low back pain which has been shown not to be associated with improved outcomes), which have flow-on effects for further low-value services (e.g., follow-up tests, referrals to specialists, procedures) which are often difficult to assess. [6,7] A total cost of care measure like TPCC is able to capture these downstream costs and minor actions from clinicians that can help curb health care costs.

The clinical logic of attributing the measure to clinicians who provide primary care services is to account for the wide range of clinicians who can provide this type of care, regardless of their designated specialty. For example, the majority of clinicians who billed Medicare as hospitalists in 2017 after the introduction of that

specialty code had previously reported a primary care specialty code in 2016.[4] Many beneficiaries see a nurse practitioner (NP) or physician assistant (PA) for their primary care, with 16 percent reporting that they saw an NP or PA for all their primary care and 29 percent saying they saw an NP or PA for some of their primary care.[4] Both primary care clinicians and specialists can provide care for an 'index condition'.[5] Multidisciplinary teams providing primary care have become more common, particularly for preventative services.[10] This underscores the need for the TPCC measure to be constructed in a way that identifies and attributes clinicians who provide primary care through the services that they provide and accounts for team-based care.

The TPCC measure accounts for patients with comorbidities through the use of different risk models. This is because patients with comorbidities are associated with higher resource use, such as through more frequent visits to the primary care clinician and specialists.[5] The increase in percentage of beneficiaries with five or more chronic conditions has been noted by MedPAC as having fueled rapid growth in Medicare spending.[4] These underscore the importance of a risk adjustment model that accounts for comorbidities, as well as interactions between comorbidities.

The measure methodology was developed with input from a technical expert panel and 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, Herbert L. "Cutting the Cost of Health Care: The Physician's Role." Texas Heart Institute Journal, vol. 43, no. 1, 2016.

[2] Crosson, FJ. "Change the microenvironment. Delivery system reform essential to control costs." Mod Healthc., vol. 39, no. 17, 2009, pp. 20-1

[3] American Academy of Family Physicians. "Valuation of Care Management Performed by Primary Care Services: An Issue Brief.", 2018

[4] MedPAC. "Report to the Congress: Medicare Payment Policy,", 2019, http://www.medpac.gov/docs/default-source/reports/mar19_medpac_entirereport_sec.pdf

[5] Starfield, B, Lemke K, Bernhardt R, Foldes S, Forrest C, and Weiner J. "Comorbidity: Implications for the Importance of Primary Care in Case Management." The Annals of Family Medicine 1, no. 1 (January 2003): 8–14. https://doi.org/10.1370/afm.1

[6] MedPAC. "Report to the Congress: Medicare and the Health Care Delivery System" June 2018

[7] Fried, J, Andrew A, Ring N, Pastel D, "Changes in Primary Care Health Utilization after Inclusion of Epidemiologic Data in Lumbar Spine MR Imaging Reports for Uncomplicated Low Back Pain" Radiology, Volume 287: number 2, May 2018

[8] Choosing Wisely Campaign, http://www.choosingwisely.org/

[9] Mafi J, Russel K, Bortz B, Dachary M, Hazel W, Fendrick M, "Low-Cost, High-Volume Health Services Contribute The Most To Unnecessary Health Spending", Health Affairs, Vol 36, no 10 (Oct 2017)

[10] Rodriguez, Hector P., William H. Rogers, Richard E. Marshall, and Dana Gelb Safran. "Multidisciplinary Primary Care Teams." Medical Care 45, no. 1 (January 2007): 19–27. https://doi.org/10.1097/01.mlr.0000241041.53804.29

S.8.3a. CLINICAL LOGIC ATTACHMENT or URL: If needed, attach <u>supplemental</u> 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)*

Measure Trigger: The start of a primary care relationship between a clinician and beneficiary and is identified by the occurrence of two Part B Physician/Supplier (Carrier) claims with particular CPT/HCPCS services billed within 90 days of each other. There are two different sets of CPT/HCPCS codes used: E&M primary care services and primary care services.

E&M primary care services are a specific set of evaluation and management codes for physician visits in the outpatient setting, physician office, nursing facility, or assisted living.

Primary care services are a broader list of services related to routine primary care and generally fall into the following categories: Durable Medical Equipment (DME) and Supplies, Electrocardiogram, Laboratory - Chemistry and Hematology, Other Diagnostic Procedures (Interview, Evaluation, and Consultation), Other Diagnostic Radiology and Related Techniques, Prophylactic Vaccinations and Inoculations, Routine Chest X-ray, Clinical Labs, Preventive Services

To trigger the measure, firstly, an initial E&M primary care service billed on Part B Physician/Supplier (Carrier) claim is identified. This E&M primary care service is not considered if it occurs during a beneficiary's stay at a Critical Access Hospital (CAH), Inpatient Facility, or Skilled Nursing Facility (SNF). Secondly, in addition to the initial E&M primary care service, at least one of the following services should be billed to confirm the candidate event:

•From any TIN within +/- 3 days: Another primary care service,

•From the same TIN within + 90 days: A second E&M primary care service OR another primary care service

End mechanisms: The risk window that opens from the time that the primary care relationship is identified as beginning, ends one year from the service date of the initial E&M primary care service.

Rationale: The triggering methodology for the TPCC measure identifies primary care relationships between a clinician and a patient by requiring at least two claims with services indicative of primary care. Requiring multiple claims within a defined, relatively short period of time avoids attribution from just a single claim (a refinement from the previous version of the TPCC measure) and ensures evidence of a sustained relationship using codes representative of overall health care evaluation and management. Exclusions are applied to further protect against potential misattribution. The specialty exclusions prevent clinicians unrelated to primary care from triggering events in a clinician group. Additionally, clinicians are excluded using their billing patterns to characterize their clinical role. This includes removing clinicians exceeding a low threshold of beneficiaries in which they are providing anesthesia, global surgery, therapeutic radiation, and/or chemotherapy.

The intent of the measure is to capture primary care relationships, which by its nature, is long-term and includes all costs in a one year long observation period to provide a broad assessment of a clinician's management of the overall health of a patient, rather than a specific condition. A longer window is able to capture the costs of downstream services that are related to the scope of primary care (e.g., preventive care). Additionally, initiating a one year long observation period from the initial claim of the paired services also ensures that attributed clinicians and clinician groups know which patients will be attributed at the time of service, allowing actionability and improvement on the measure.

The trigger mechanism allows for multiple attribution of the eligible clinicians and clinician groups that are responsible for a patient's primary care management to be concurrently attributed that patient's beneficiary-months. Holding multiple clinician groups that demonstrate responsibility for the patient is not only fairer, it

encourages coordination of care between these providers. Overall, both patients and clinicians benefit when all providers involved in the care of the beneficiary are covered by similar incentives.

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 of the different CMS-HCC and CMS-ESRD models. These models include variables indicating a patient's health status at the start of each beneficiary month, allowing for the measure to capture changes in health status over time. The range of models reflect different clinical severity levels; for example, patients with ESRD have different clinical profiles from patients without ESRD. See Sections S.8.2 and S.9.3 for further details.

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 accounted for within the risk adjustment models. Where beneficiaries have sufficient Medicare medical history, the models use HCCs which are indicator variables for comorbidities and clinical conditions. As the relationship between comorbidities' resource use may be non-linear in some cases, the models take into account sets of interactions between HCCs, demographic, and/or enrollment status variables.

The CMS-HCC and CMS-ESRD models were selected based on previous studies evaluating its appropriateness for use in risk adjusting Medicare claims data. These models were 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 to predict annual cost. In addition, the CMS-HCC and CMS-ESRD models are 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. 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 models exist for use in the Part C Medicare Advantage program, testing results for factors included in the CMS-HCC V22 and CMS-ESRD V21 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 (CMS 2018).

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 TPCC measure consists of Medicare beneficiaries enrolled in Medicare Parts A and B for whom the measure identifies as having a primary care relationship with a clinician. To be included, the beneficiary must have at one of his or her beneficiary month occurring during the performance period.

Exclusions:

Several steps in the construction of the TPCC measure ensure comparability by fostering comparability in the beneficiary population captured and clinician population measured. These are detailed in Section S.7.2.

In keeping with the measure intent to capture the overall costs of care for beneficiaries receiving primary care services, there are a limited set of exclusions primarily to ensure that, as part of data processing, sufficient data are available to accurately determine resource use and calculate risk adjustment for each beneficiary. These exclusions, along with their rationales, are listed below:

•The beneficiary was not continuously enrolled in Medicare Parts A and B unless partial enrollment was the result of either new enrollment or death only. These beneficiaries may have gaps in their Medicare claim records when benefits are covered by other payers.

•The beneficiary resides outside the United States or its territories during the performance period. Differences in reimbursement policy for healthcare services provided outside the U.S. can lead to unfair comparisons of cost.

•The beneficiary receives benefits from the Railroad Retirement Board (RRB). Beneficiaries covered by the RRB may have healthcare benefits normally covered by Medicare paid by the RRB, which may bias the observed cost for these beneficiaries.

To ensure the clinicians attributed the measure are within the intended scope of primary care management, exclusions of clinicians are used to ensure comparability. Clinicians who would not reasonably be responsible for providing primary care are excluded from attribution of the revised TPCC measure using their CMS HCFA specialty designation assigned on Part B physician/supplier claims. This exclusion aims to keep primary care specialists and internal medicine subspecialists who frequently manage patients with chronic conditions falling in their areas of specialty. Additionally, clinicians are characterized by their Part B billing behavior and excluded from attribution if found meeting a threshold of billing for the following service categories; 10-day or 90-day global surgery services, anesthesia services, therapeutic radiation services, chemotherapy services. The methodology and clinical logic for exclusions of clinicians from attribution is further detailed in Section S.8.2

Data truncation is applied to risk-adjusted beneficiary monthly costs for outlier values through winsorization on the right tail. Monthly costs at the 99th percentile are assigned to all attributable beneficiary months with costs above the 99th percentile. Winsorization aims to limit the effects of extreme values on expected costs. Winsorization is a statistical transformation that limits extreme values in data to reduce the effect of possible outliers. The risk adjustment approach is detailed in Section S.7.2 and in S.9.3.

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)*

Differences in patient case mix are accounted for by using separate risk adjustment models for the following types of beneficiaries, as discussed in Section S.7.2:

1) Beneficiaries without ESRD

1a) Beneficiaries with fewer than 12 months of Medicare medical history

2a) Beneficiaries with at least 12 months of Medicare medical history

3a) Beneficiaries in long-term institutional care settings

2) Beneficiaries with ESRD receiving dialysis

2a) Beneficiaries with fewer than 12 months of Medicare medical history

2b) Beneficiaries with at least 12 months of Medicare medical history

This stratification accounts for the very different patient clinical profiles for patients with ESRD receiving dialysis and patients without ESRD, as well as maximizes the availability of Medicare claims history to be able to construct indicator variables for clinical conditions.

The TPCC measure uses the CMS-HCC V22 risk adjustment models for new enrollee, community, and long-term institutional beneficiaries without ESRD. A beneficiary month is measured under the new enrollee model if they do not have a full one-year lookback of Medicare claims data as of the start of a beneficiary month. As a

result, the model is derived primarily from beneficiary enrollment data. This model adjusts for gender, age, dual Medicare and Medicaid enrollment, and whether the beneficiary was originally entitled to Medicare due to disability through a series of interacted covariates. Beneficiaries with sufficient Medicare claims history are measured under the community or the institutional model if they are institutionalized in a long term care facility. In both models, severity of illness is measured using HCCs and disease interactions. 79 HCCs are accounted for under CMS-HCC V22 model for beneficiaries classified as community enrollees and long-term institutional enrollees while the exact number and types of disease interaction can vary. Both models interact beneficiary age with gender. In addition, the community model interacts dual enrollment status, gender, and the indicator for whether the beneficiary was originally entitled to Medicare due to disability, while the institutional model adjusts for disability as the original reason for Medicare enrollment and dual enrollment status independently.

For ESRD beneficiaries receiving dialysis, the TPCC measure utilizes the CMS-ESRD V21 risk adjustment models. Differentiated models are implemented for dialysis new enrollees and dialysis community enrollees. Similar to the CMS-HCC V22, enrollees are classified as new enrollees if they were not continuously enrolled in Parts A and B for the one-year lookback period prior to each beneficiary month. As a result of this, the model primarily uses information from the beneficiary's enrollment data. This model adjusts for gender, age, dual enrollment status, and whether the beneficiary was originally entitled to Medicare due to disability through a series of interacted covariates. In addition to accounting for these patient characteristics, the dialysis community model also risk adjusts for medical severity using 87 HCCs and additional disease interactions.

The CMS-ESRD V21 and CMS-HCC V22 models both generate a risk score for each beneficiary that summarizes the beneficiary's expected cost of care relative to other beneficiaries. Risk scores for ESRD beneficiaries are normalized to enable comparison with the HCC V22 risk scores. This is achieved by multiplying ESRD risk scores by the mean annual Medicare spending for the ESRD population applied in the CMS-ESRD V21 model and dividing by the mean annual Medicare spending for the total Medicare population applied in the CMS-HCC V22 model, effectively renormalizing ESRD risk score values to the equivalent scale of the HCC models. A risk score equal to one indicates risk associated with expenditures for the average beneficiary nationwide. Risk scores below or above one indicate below and above average risk, respectively.

The complete list of risk adjustment variables for each model are listed in the Measure Codes List linked in Section S.1 in the tab titled HCC_Risk_Adjust.

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):*

Continuous variable

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 TPCC measure score is the average payment-standardized, risk-adjusted, and specialty-adjusted monthly cost across all beneficiary months in the performance period attributed to a clinician or 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 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 included in the calculation.

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

As described in Section S.7.2, the TPCC measure is calculated for each clinician and clinician group practice by averaging the risk-adjusted and specialty-adjusted cost across the beneficiary months attributed. Adjustments to observed monthly costs are calculated as follows:

1) Divide observed costs for each beneficiary month by the normalized risk score to obtain risk-adjusted monthly costs.

2) Winsorize risk-adjusted monthly costs at the 99th percentile by assigning the 99th percentile of monthly costs to all attributable beneficiary months with costs above the 99th percentile.

3) Normalize monthly costs to account for differences in expected costs based on the number of clinician groups to which a beneficiary is attributed in a given month. The normalization factor is the inverse cube root of the number of attributed clinician groups for that beneficiary month.

4) Calculate the average risk-adjusted monthly cost for each TIN and TIN-NPI by averaging risk-adjusted monthly cost across all attributed beneficiary months.

5) Calculate the national specialty-specific expected cost for each specialty as the weighted average of TIN/TIN-NPI's risk-adjusted monthly cost.

5a) Define the weight for each TIN/TIN-NPI as the percentage of clinicians with that specialty multiplied by the total number of beneficiary months attributed to the TIN/TIN-NPI multiplied by the number of clinicians with that specialty.

6) Calculate the specialty-adjustment factor for each TIN or TIN-NPI as follows:

6a) Multiply the national specialty-specific expected cost for each specialty by the respective specialty's share of Part B payment within a TIN or TIN-NPI and sum the weighted share of national specialty-specific expected cost calculated in the previous step across all the specialties under a given TIN or TIN-NPI.

7) Calculate final risk-adjusted, specialty-adjusted cost measure by dividing each TIN and TIN-NPI's average risk-adjusted monthly cost by their specialty-adjustment factor and multiply this ratio by the average non-risk-adjusted, winsorized observed cost across the total population of attributed beneficiary months.

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 TPCC 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 TPCC 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 over 550,000 field test reports containing cost measure performance on the draft TPCC measure as specified at that time were available to clinicians and clinician groups meeting a 20-beneficiary case minimum (120,266 TIN-level reports and 446,973 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 TPCC measure is attributed to a TIN billing two Part B Physician/Supplier (Carrier) claims with particular CPT/HCPCS services billed within 90 days. There are two different sets of CPT/HCPCS codes used: E&M primary care services and primary care services. E&M primary care services are a specific set of evaluation and management codes for physician visits in the outpatient setting, physician office, nursing facility, or assisted living. Primary care services are a broader list of services related to routine primary care. The pairing of these code sets is used because they represent primary care services and requires more than one claim to confirm that a clinical relationship has been established.

Once clinician exclusions are applied, the individual clinician (TIN-NPI) within a TIN that provides the most primary care evaluation and management services for the beneficiary is attributed their respective attribution events.

Based on input from a technical expert panel, this attribution methodology was incorporated as a refinement as part of the comprehensive re-evaluation process that the version of the TPCC measures (used in MIPS 2017-19) underwent. This revised methodology accounts for the nature of primary care where multiple clinicians can have an ongoing relationship with a beneficiary. This attribution methodology also prevents attribution of a beneficiary prior to a clinician meeting him or her, which was part of the previous version of the measure.

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 primary care services to beneficiaries, as identified by meeting the logic described in Section S.7.2 and S.13.2 to identify when a primary care relationship has begun. The peer group is limited to clinicians who are reasonably providing primary care. This is achieved by excluding clinicians who are unlikely to be providing primary care either based on HCFA specialty designation or clinician billing patterns. The rationale for identifying the peer group in this way is to focus the measure on clinicians providing primary care, in line with the intent of the measure and 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, such as to adjust for the specialty composition of a clinician group.

S.13.4. Sample size

Detail the sample size requirements for reporting measure results.

From the MIPS CY 2020 performance year and onwards, the TPCC measure will be calculated and reported via confidential reports for TINs and TIN-NPIS with 20 or more attributed beneficiaries. 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 the average payment-standardized, risk-adjusted, and specialty-adjusted monthly costs across all beneficiary months in the performance period attributed to a TIN or TIN-NPI. 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 primary care management to hold clinicians accountable for their cost effectiveness. Combined with measures in the other MIPS performance categories, such as the quality performance category, the TPCC 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-tpcc-v6.docx

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

Measure Number (*if previously endorsed*)**:** N/A
Measure Title: Total Per Capita Cost (TPCC)
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 TPCC measure uses Medicare Part A and Part B claims data maintained by CMS. These claims data are used to attribute beneficiary months, calculate beneficiary's monthly costs, and construct risk adjustors. Data from the EDB are 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 models also account for expected differences in payment for services provided to beneficiaries in long-term care based on the data from the MDS. Specifically, the MDS is used to identify beneficiaries that should be risk adjusted through the CMS-HCC v22 institutional model.

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? The measurement period for the TPCC testing is January 1, 2018 through December 31st, 2018. The split-sample intraclass correlation also includes data from the 2017 fiscal year. Attribution validity in Section 2b1.3 also uses the 2017 fiscal year. For further details, please see Question 1.7.

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> 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*)

74,191 clinician group practices (identified by Tax Identification Number [TIN]) and 335,480 practitioners (identified by combination of TIN and National Provider Identification [NPI]) were included in the analyses. Clinicians and clinician groups were included in testing if they were attributed 20 or more TPCC beneficiaries during the measurement period of calendar year 2018. Beneficiaries from all 50 States and D.C. receiving evaluation and management care indicative of primary care were included, with their respective costs evaluated from all claim settings.

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*)

26,636,602 Medicare beneficiaries (with 305,918,850 beneficiary months) were included in TIN level testing and analysis, and 26,374,993 beneficiaries (with 297,584,563 beneficiary months) included in TIN-NPI level measure testing.

The beneficiary population eligible for the TPCC attribution consists of Medicare beneficiaries enrolled in Medicare Parts A and B (but not Part C) receiving evaluation and management services that indicate a primary care relationship. Beneficiaries were included in the sample if they met a set of inclusion criteria meant to ensure completeness of data. The inclusion criteria are:

- The beneficiary has Medicare as their primary payer for the entire measurement period.
- The beneficiary was continuously enrolled in Medicare Parts A and B and any instance of partial enrollment was the result of either new enrollment or death only.
- The beneficiary date of birth is not missing.
- The beneficiary did not reside outside the United States or its territories during any month of the measurement period
- The beneficiary is not covered by the Railroad Retirement Board

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

This analysis shows that the TPCC measure's inclusion criteria have only a minimal effect on the percentage of total beneficiaries of any particular demographic at the TIN level (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 -2.5 and 2.3 percentage points. To illustrate, the percentage of beneficiaries aged 65 to 69 without

applying the inclusion criteria is 30.8 percent, compared to 31.2 percent with the inclusion criteria at TIN level testing. The difference in the proportion of male and female beneficiaries is less than 0.01 percentage points when comparing the application of inclusion criteria. The percentage of beneficiaries identified as female is 56.2 percent with or without applying inclusion criteria. 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 data from calendar years 2018 and fiscal year 2017. Attribution validation (Section 2b1.3, Tables 5 and 6) was conducted on fiscal year 2017 data. 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. Social risk variables analyzed from the EDB and CME were sex and dual status.The ACS was used to calculate the Agency of Healthcare Research and Quality (AHRQ) SES Index. The AHRQ index scores are calculated using the AHRQ scoring algorithm and is a continous dependent variable as a replacement of all of the followign 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.[1,2]

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*)

---

[1] 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

[2] 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])

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 scores are 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 clinicians within the measure. 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 beneficiaries from an aggregation of two years of data (fiscal year 2017 and calendar year 2018). We used two years of data to achieve numbers of beneficiaries 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 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 meet a case minimum of 20 beneficiaries in both performance periods were included. The sample was stratified by the performance years, thus ensuring that beneficiaries within each 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 estimated 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 20 beneficiaries (the case minimum for the measure in the MIPS 2020 performance period) the mean reliability for TINs is 0.84 and for TIN-NPIs is 0.88. 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.[3] Mean reliability increases with increasing

[3] 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

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 Score Results for TINs and TIN-NPIs with an Overall Testing Volume Threshold of 20 Beneficiaries

Reporting Level	Number of TINs or TIN-NPIs	Mean (Std. Dev.)	25th Pct.	50th Pct.	75th Pct.
TIN	74,191	0.84 (0.14)	0.77	0.89	0.95
TIN-NPI	335,480	0.88 (0.08)	0.83	0.91	0.95

* 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 volume of clinicians within the practice, the average reliability scores increases from 0.81 (1 clinician) to 0.94 (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	**74,191**	**0.84 (0.14)**	**0.77**	**0.89**	**0.95**
1 Clinician	37,657	0.81 (0.13)	0.73	0.85	0.91
2-4 Clinicians	17,042	0.86 (0.13)	0.80	0.91	0.95
5-20 Clinicians	12,842	0.89 (0.13)	0.85	0.94	0.97
21+ Clinicians	6,650	0.94 (0.11)	0.93	0.98	0.99

* Pct. = percentile.

Split-sample Reliability Testing. Table 3 presents ICC(2,1) between the split-sample measure scores for the overall sample of 68,413 TINs and 265,106 TIN-NPIs included in this testing. The ICC in the overall sample for TIN reporting was 0.76 and 0.64 for TIN-NPI reporting.

Table 3. Split-sample Intraclass Correlation Coefficients

Reporting Level	# of TINs or TIN-NPIs	Mean Score FY 2017	Mean Score CY 2018	Pearson Correlation Coefficient	ICC(2,1)
TIN	68,413	$1,089	$1,089	0.76	0.76

Reporting Level	# of TINs or TIN-NPIs	Mean Score FY 2017	Mean Score CY 2018	Pearson Correlation Coefficient	ICC(2,1)
TIN-NPI	265,106	$1,143	$1,143	0.64	0.64

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 0.84 for TINs and 0.88 for TIN-NPIs at a volume threshold of 20 beneficiaries[4]. Reliability for groups of different practice sizes was similar, with mean reliability for the smallest TINs at 0.81.

The split-sample reliability analysis provides further evidence of reliability and repeatability of the performance measure. Reliability (ICC(2,1)) was 0.76 for TINs and 0.64 for TIN-NPIs, which indicates good or high 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 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**

---

[4] 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).

☒ **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 TPCC measure underwent a structured process for gathering detailed input from experts and a broad range of stakeholders. An expert panel was convened to provide input to inform the comprehensive measure re-evaluation process. Experts provided input on measure refinements (e.g., attribution logic) that would help ensure that the measure is fulfilling its intent to capture the overall costs of care for patients as a result of their primary care management (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, and (ii) 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. At a TEP meeting in August 2017, the panel provided high-level guidance and initial input on direction for refinement, focusing on which clinician(s) to attribute and the timing of attribution 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 primary care management, which is key to measure actionability and meaningfully differentiating between providers. During a May 2018, the TEP provided further input on specific approaches with empirically testing to refine the attribution methodology. Based on input from the first two TEP meetings, TPCC attribution rules were refined to better identify care relationships and fairly attribute clinicians using a pairing of primary care related claims to address previous stakeholder comments. For example, the timing of attribution was refined so that cost could no longer be assigned prior to seeing a beneficiary, and clinicians were excluded based on their Part B billing patterns.

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, 567,239 field test reports for TINs and TIN-NPIs were available for download and review for the TPCC measure revised in 2018. Following field testing and with the input from the TEP who met in November 2018 to consider field testing feedback, a number of refinements were made to the measure including adding a specialty adjustment to account for differences in TIN's specialty composition and creating specialty exclusions to remove from attribution those clinicians belonging to specialties that are unlikely to be responsible for primary care.

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 TPCC measure captures variation in service utilization by examining differences in mean risk- and specialty-adjusted cost for beneficiary months stratified by beneficiaries with known indicators of resource or service utilization, specifically complications related to acute admission and post-acute care utilization. For this analysis, we compared the mean risk- and specialty-adjusted monthly cost for beneficiaries with and without complications related to acute admission and post-acute care utilization occurring in the measurement period. We expected that beneficiaries as measured by TPCC with these indicators of resource or service utilization would be more expensive those 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- and specialty-adjusted monthly costs. We classified certain services included in the TPCC measure into clinically coherent groups of services, called "clinical themes". The clinical themes are:

- **Acute Inpatient Service,** including acute inpatient hospitalization and the related services billed by any clinician during the stay
- **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 service category and the risk- and specialty-adjusted cost.

We hypothesized that at least some of the Post-Acute Care categories and the Acute Inpatient Services category would have the highest correlation with risk- and specialty-adjusted beneficiary monthly cost even after accounting for beneficiary characteristics, as these types of care are often associated with costly services related to treatment of complications.

We also examined the possibility of testing a hypothesized relationship between clinicians' TPCC scores and their scores on MIPS quality measures. This type of testing assesses whether clinicians with better TPCC 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 TPCC 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 primary care and general health outcomes that would be reflected in the cost measure. Of the available outcome measures, some were aimed at different types of care from what TPCC aims to cover (e.g., MIPS Q#342 is specific to admission to palliative care services) and others were narrowly specified relative to the intent of the TPCC measure (e.g., MIPS Q#338 applies only to patients with a diagnosis of HIV). MIPS Q#398 Optimal Asthma Control (eCQM measure) has potential to explore a conceptual relationship with the TPCC measure. The quality measure assesses two components of asthma control: performance on asthma control tests/questionnaires, and whether the patient was at risk of exacerbation, defined as having less than 2 ED visits and inpatient hospitalizations due to asthma in 12 months. While this suggests that a conceptual relationship as ED visits and inpatient hospitalizations will be reflected in claims data, we note the following critical limitations: (i) the measure is specified for patients aged 5-17 and 18-50, which greatly reduces the common patient cohort with TPCC, and (ii) only 376 MIPS participating clinicians or groups reported this measure in 2017.

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.[5] The measure construction also means that even with complete data, the patient cohort and attributed clinicians would differ; for example, MIPS Q#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. Intermediate outcome measures (e.g., MIPS Q#236 Controlling High Blood Pressure) have a weaker conceptual relationship with a cost measure, as the link between one point in time blood pressure test and observable costs in claims data is much more remote than an outcomes measure. As such, we were unable to test the conceptual relationship between the TPCC measure and related quality measures.

Attribution Validity Testing: As part of earlier testing on the measure specifications, we also evaluated the validity of the TPCC measure attribution methodology through two approaches. Firstly, we examined the proportion of the beneficiary's Part B Evaluation and Management (E&M) codes related to primary care that are billed by the attributed TIN/TIN-NPI, to demonstrate that there is claims-based evidence that those TIN/TIN-NPIs manage their beneficiaries' ongoing care. And secondly, we conducted an impact analysis on the volume of TINs attributed the measure solely based on the services conducted by their Nurse Practitioners (NP) and/or Physician Assistants (PA), to check if TINs unlikely to manage primary care are attributed through the work of the NP and PA within their practice. These tests are aimed at evaluating the extent to which the TPCC measure is attributing beneficiaries to clinicians who are providing primary care services and who have a relationship with a given beneficiary.

We expected that for the first approach, there would be a high level of engagement between attributed clinicians and beneficiaries, as measured by a higher proportion of the beneficiary's E&M codes billed by the attributed clinicians. For the second approach, we expected limited attribution of TINs with specialties unlikely to manage primary care via their NPs and PAs alone.

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

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

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 the key refinements help the measure provide an accurate reflection of the overall costs related to primary care: (i) Identifies the primary care relationship by accounting for the overall pattern of primary care service delivery, and (ii) Allows for attribution of costs to multiple clinicians and clinician groups

Response: 12 members agreed (rating between 4-6), 3 members disagreed (rating between 1-3)
Mean Rating[6]: 4.9 out of 6 (somewhat to moderate agreement)

Question 2: Indicate the extent to which you agree with the following statement comparing the refined TPCC 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 TPCC measure provide a more accurate reflection of the costs for overall primary care than the previous version of the measure, and can better distinguish good and poor performance."

Response: 13 members agreed (rating between 4-6), 2 members disagreed (rating between 1-3)
Mean Rating: 4.8 out of 6 (somewhat to moderate agreement)

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

Response: 12 members agreed (rating between 4-6), 3 members disagreed (rating between 1-3)
Mean Rating: 4.8 out of 6 (somewhat to moderate agreement)

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: Two members who disagreed with the statement in Question 3 provided comments. These related to concerns about double counting costs with episode-based cost measures. Another member noted that while the refinements are an improvement on the measure, the attribution methodology should identify one PCP (e.g., through the use of a primary care add-on code).

Other members who agreed with the statement provided general feedback in this question, such as recommending further review across different measures.

Empirically Validity

Table 4 present results for the first analysis of validity that shows the distribution of beneficiary's mean risk- and specialty-adjusted monthly cost across beneficiary months for a beneficiary during the measurement period. The mean of beneficiary's average risk- and specialty-adjusted monthly cost for a beneficiary during the measurement period is $1,187. The mean of beneficiary's average risk- and specialty-adjusted monthly cost for beneficiaries with services relating to acute inpatient admissions is $2,647, compared with $866 for a beneficiary without acute inpatient admissions. The mean of beneficiary's average risk- and specialty-adjusted monthly cost with services relating to Post-Acute Care is $2,427, compared with $996 for a beneficiary without PAC.

Table 4. Distribution of Beneficiary's Average Risk- and Specialty-Adjusted Monthly Cost

[6] 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	Beneficiary Mean Risk- and Specialty-Adjusted Monthly Cost						
	Mean	Std. Dev.	Percentiles				
			10th	25th	50th	75th	90th
All Beneficiaries	$1,187	$1,567	$148	$302	$669	$1,509	$2,758
Beneficiaries with Acute Inpatient Admissions	$2,647	$2,211	$882	$1,366	$2,119	$3,175	$4,761
Beneficiaries without Acute Inpatient Admissions	$866	$1,161	$128	$255	$516	$1,035	$1,948
Beneficiaries with Post-Acute Care (IRF, LTCH, HH, SNF)	$2,427	$2,048	$650	$1,140	$1,969	$3,055	$4,552
Beneficiaries without Post-Acute Care (IRF, LTCH, HH, SNF)	$996	$1,383	$134	$269	$564	$1,201	$2,283

Table 5 includes results from the clinical theme analysis. The service categories analysis shows the correlation between cost observed in a clinical theme and risk- and specialty-adjusted cost at the TIN and TIN-NPI levels. At both the TIN and TIN-NPI levels, there is a strong correlation between the SNF service category and risk-adjusted cost (correlation: 0.54). At both the TIN and TIN-NPI levels there is a strong correlation between Outpatient E&M Services, Procedures, and Therapy and risk-adjusted cost (correlation: 0.45). At both the TIN and TIN-NPI levels, there is a moderate to large correlation between the Acute Inpatient Services category and risk-adjusted cost (correlation: 0.38). In contrast, at the TIN and TIN-NPI levels the HH category has a lower correlation with risk-adjusted cost (correlation: 0.11). Similarly, there is lower correlation between the Non-Hospital Admission Emergency Services category and risk-adjusted cost (correlation: 0.15). [7]

Table 5. Pearson Correlation Statistics between Costs of Clinical Themes with Risk-Adjusted Cost

Clinical Theme	Pearson Correlation	
	TIN	TIN NPI
Acute Inpatient Services	0.38	0.38
Emergency Services Not Included in Hospital Admission	0.15	0.15
Outpatient E&M Services, Procedures, and Therapy	0.45	0.45
Post-Acute Care: Home Health	0.11	0.11
Post-Acute Care: IRF/LTCH	0.18	0.18
Post-Acute Care: SNF	0.54	0.54

[7] 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.

Attribution Validity

The first analysis on attribution validity (**Table 6**) shows the distribution of the share of beneficiary's primary care related E&M claims billed by the attributed TIN or TIN-NPI. The mean share of beneficiary's E&M claims billed by attributed TINs or TIN-NPI is 52.8 percent and 45.0 percent, respectively.

Table 6. Share of Primary Care E&M Billed by Attributed TIN and TIN-NPI

Provider Type	Distribution of Share of Attributed Beneficiary's Primary Care Related E&Ms Billed by Attributed TIN or TIN NPI			
	Mean	25th Percentile	Median	75th Percentile
TIN	52.8%	34.8%	55.8%	70.0%
TIN-NPI	45.0%	28.9%	42.8%	61.6%

The second analysis on attribution validity (**Table 7**) shows the share of TINs attributed the measure via NPs and/or PAs alone. Only 13.3 percent of all TINs were attributed based on services conducted by NPs and/or PAs exclusively; where 7.8 percent of this total come from TINs comprised of a majority of NP and/or PAs. For TINs with specialties not primary consisting of NP or PA, only 5.5 percent of all TINs are attributed via this method.

Table 7. Frequency of Most Common HCFA Specialties in TINs Attributed TPCC Measure via Nurse Practitioners and Physician Assistants Alone

Specialty	% of All Attributed TINs
All TINs Attributed via NP/PA Alone	**13.3%**
Nurse Practitioner	6.7%
Physician Assistant	1.1%
Orthopedic Surgery	1.1%
Psychiatry	1.0%
Otolaryngology	0.4%
Urology	0.4%
Neurology	0.3%
Physical Medicine and Rehabilitation	0.3%
General Surgery	0.2%
Interventional Pain Management	0.2%
All Other Specialties	1.6%

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, 12 (80%) 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 risk- and specialty-adjusted monthly costs for beneficiaries with acute inpatient admissions and post-acute care in the measurement period are higher than for beneficiaries without those services. This indicates that the measure can capture higher resource use by clinicians who have higher rates of complications related to these types of services, while not disincentivizing the provision of appropriate care in other areas.

The second test, the clinical themes analysis, demonstrates that the TPCC measure is able to accurately capture higher resource use across various types of services. Importantly, we see that the correlation with risk- and specialty-adjusted cost is strong not only for high-cost categories such as Acute Inpatient Services (average monthly cost for TIN testing $9,670 and $9,664 for TIN-NPI testing), but also for lower cost categories such as outpatient E&M services, procedures, and therapy (average monthly cost for TIN testing $712 and $713 for TIN-NPI testing). This indicates that the risk- and specialty-adjusted cost increase is not just a result from a mechanical increase in beneficiary month costs from high-cost categories but also can respond to over utilization of lower cost clinical theme. 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.

Finally, testing on attribution validity showed that the measure is appropriately identifying and attributing beneficiaries to clinicians who have a primary care relationship with them. Attributed TIN/TIN-NPIs bill a large proportion of beneficiaries' E&M claims related primary care as was intended with the re-evaluation effort for the attribution methodology. This indicates that there is a strong relationship between attributed TIN/TIN-NPIs and beneficiaries that they treat. The results of the analysis for attribution as a result of NP/PA show that the TPCC measures attributes few TIN unlikely to provider primary care through NPs and PAs alone.

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*)

TPCC aims to measure a broad cohort of Medicare patients who receive primary care services. Exclusions are primarily used to ensure that, as part of data processing, sufficient data are available to accurately determine resource use and calculate risk adjustment for each beneficiary. The exclusions focus on removing beneficiaries where fair comparisons cannot be made across providers. These exclusions, along with their rationales, are listed below:

- The beneficiary was not continuously enrolled in Medicare Parts A and B unless partial enrollment was the result of either new enrollment or death only.
    - These beneficiaries may have gaps in their Medicare claim records when benefits are covered by other payers.
- The beneficiary resides outside the United States or its territories during the measurement period
    - We may not have complete claims data available for beneficiaries outside the US and territories.
- The beneficiary receives benefits from the Railroad Retirement Board (RRB)
    - Beneficiaries covered by the RRB may have healthcare benefits normally covered by Medicare paid by the RRB, which may bias the observed cost for these beneficiaries.

Given the rationales for these exclusions, we would expect these excluded beneficiaries' Medicare costs to have different rates and measurability than the included beneficiaries. For the exclusions, we examined annual Medicare Parts A and B spending from potentially attributable triggering events (i.e., candidate events) for excluded beneficiaries compared to spending for beneficiaries included in measure calculation to assess the differences 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 8 below presents observed annual cost statistics from candidate events for beneficiaries who are excluded from the TPCC measure and from the set of final candidate events for beneficiaries included in the TPCC measure. Appendix Table 2b2.2 provides more detailed cost distributions for the excluded populations.

Table 8. Cost Statistics for Measure Exclusions

Exclusion	Beneficiaries		Observed Cost		
			Mean	Percentile	
	#	%		10th	90th
Not Continuous Enrollment In Medicare Part A and B or Any Enrollment in Part C	5,053,189	14.33%	$19,030	$461	$52,463
Beneficiary Resides Outside of U.S. or Territories	14,879	0.04%	$11,123	$253	$31,305
Beneficiary Enrollment in Medicare for Railroad Workers and their Families	326,588	0.93%	$17,585	$952	$48,245
Remaining Candidate Events after Beneficiary-level Exclusions	29,921,474	84.86%	$18,345	$1,065	$49,655

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 mean observed cost of candidate events for beneficiaries without continuous enrollment in Medicare Parts A and B is slightly higher compared to the mean observed cost for final candidate events. Though at the 10th percentile of the distribution of the excluded beneficiaries have much lower annual spending than compared to the final population of beneficiaries. These results indicate that including these beneficiaries observed annual cost cannot be fairly compared to continuously enrolled Part A and B beneficiaries. Additionally, implementing this exclusion mitigates the extent to which gaps in a beneficiary's claims history can adversely affect the determination of risk factors for risk adjustment. For example, certain conditions that arise and should be reflected in the risk adjustment model may not be observed in the available claims data due to gaps in the beneficiary's Part A and B claims history.

Cases where the beneficiary resides outside the United States or its territories during the measurement period are associated with even lower costs. This difference from the final candidate events is pronounced at both ends of the distribution. This could be due to different care systems and reimbursement policies outside of the U.S. and its territories. Exclusion of these beneficiaries is therefore appropriate to ensure that we have complete claims data.

Beneficiaries that receive benefits from the RRB also have slightly lower mean cost compared to final candidate events after the exclusions are applied. While costs for candidate events for these beneficiaries do not substantially differ from the final set of candidate events, their Medicare claims history might have gaps in cases where their services are paid for by the RRB.

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** 28-133 **risk factors**

☒ **Stratification by** 5 **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 patient case mix are controlled for using separate CMS Hierarchical Condition Category Version 22 (CMS-HCC V22) models for new enrollees, continuing enrollees, enrollees in long-term institutional settings. In addition, for beneficiaries with ESRD, the CMS ESRD Version 21 (CMS-ESRD V21) models are used for new enrollees with ESRD, and community enrollees with ESRD. The CMS models were developed for use in the Medicare Advantage program and the accuracy of the continual upkeep and performance of these models is reported to Congress every three years under the 21st century Cures Act.[8] The model is not re-predicted for TPCC, but instead applies the coefficients from the Medicare Advantage models.

The TPCC measure follows the CMS-HCC V22 risk adjustment models for new enrollee, community, and long-term institutional beneficiaries without ESRD. A beneficiary month is measured under the new enrollee model if they do not have a full one-year lookback of Medicare claims data as of the start of a beneficiary month. As a result, the model is derived primarily from beneficiary enrollment data. This model adjusts for gender, age, dual

[8] CMS, "Report to Congress: Risk Adjustment in Medicare Advantage December 2018," https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/RTC-Dec2018.pdf

Medicare and Medicaid enrollment, and whether the beneficiary was originally entitled to Medicare due to disability through a series of interacted parameters. Beneficiaries with sufficient Medicare claims history are measured under the community or the institutional model if they are institutionalized in a long term care facility. In both models, severity of illness is measured using HCCs and disease interactions. 79 HCCs are accounted for under CMS-HCC V22 model for beneficiaries classified as community enrollees and long-term institutional enrollees while the exact number and types of disease interaction can vary. Both models interact beneficiary age with gender. In addition, the community model interacts dual enrollment status, gender, and the indicator for whether the beneficiary was originally entitled to Medicare due to disability, while the institutional model adjusts for disability as the original reason for Medicare enrollment and dual enrollment status independently.

For ESRD beneficiaries receiving dialysis, the TPCC measure uses the CMS-ESRD V21 risk adjustment models. Differentiated models are implemented for dialysis new enrollees and dialysis community enrollees. Similar to the CMS-HCC V22, enrollees are classified as new enrollees if they were not continuously enrolled in Parts A and B for the one-year lookback period prior to each beneficiary month. As a result of this, the model primarily uses information from the beneficiary's enrollment data. This model adjusts for gender, age, dual enrollment status, and whether the beneficiary was originally entitled to Medicare due to disability through a series of interacted parameters. In addition to accounting for these patient characteristics, the dialysis community model also risk adjusts for medical severity using 87 HCCs and additional disease interactions.

The CMS-ESRD V21 and CMS-HCC V22 models both generate a risk score for each beneficiary that summarizes the beneficiary's expected cost of care relative to other beneficiaries. Risk scores for ESRD beneficiaries are normalized to enable comparison with the HCC V22 risk scores. This is achieved by multiplying ESRD risk scores by the mean annual Medicare spending for the ESRD population applied in the CMS-ESRD V21 model and dividing by the mean annual Medicare spending for the total Medicare population applied in the CMS-HCC V22 model, effectively renormalizing ESRD risk score values to the equivalent scale of the HCC models. A risk score equal to one indicates risk associated with expenditures for the average beneficiary nationwide. Risk scores below or above one indicate below and above average risk, respectively.

Following the normalization of risk scores, observed costs for each beneficiary month are divided by the normalized risk score to obtain risk-adjusted monthly costs. These costs are then winsorized at the 99th percentile by assigning the 99th percentile of monthly costs to all attributed beneficiary months with costs above the 99th percentile. Finally, monthly costs are normalized to account for differences in expected costs based on the number of clinician groups to which a beneficiary is attributed in a given month. This normalization is applied by dividing monthly costs by the cube root of the number of TINs to which a beneficiary is attributed for a particular a month.

Full details of the risk adjustment model are in the Measure Codes List File (linked in Section S.1).

2b3.2. If an outcome or resource use component measure is <u>not risk adjusted or stratified</u>, provide <u>rationale and analyses</u> 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 <u>and</u> 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: We selected the CMS-HCC V22 and CMS-ESRD V21 models based on previous studies evaluating their appropriateness for use in risk adjusting Medicare claims data. These models were developed specifically for use in the Medicare population, meaning that they account for conditions found in the Medicare population and are calibrated on Medicare fee-for-service beneficiaries. In addition, the CMS-HCC and ESRD models are routinely updated for changes in coding practices (e.g., the transition from ICD-9 to ICD-10 codes) and they are exhaustive on these code sets. Because the CMS-HCC and ESRD models have already been extensively tested and are used for a large Medicare Part C population, we reference this testing from the

December 2018 CMS Report to Congress on Risk Adjustment in Medicare Advantage and focus any additional testing on how the CMS-HCC and ESRD models influence the final TPCC measure score.[9]

Social Risk Factors: According to a 2014 National Quality Forum report[10], 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.[11,12,13] Testing for social risk factors is further detailed in section 2b3.4b.

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

☒ **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). The risk model in TPCC applies coefficients (i.e. risk scores) as predicted by the CMS-HCC and ESRD-HCC models in Medicare Advantage. Testing results for factors included in the CMS-HCC V22 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.[14,15]

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*

---

[9] In 2018, 20 million beneficiaries were enrolled in Medicare Part C plans and incurred $230 billion of Medicare Part A and Part B costs (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

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

[11] 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

[12] 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

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

[14] 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.

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

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.**

The CMS-HCC V22 and CMS-ESRD V21 include risk adjustors for gender and low income status (as identified through dual enrollment status) in a series of interacted variables to account for social risk factors. The Medicare Advantage program uses these risk adjustment models to predict total annual cost of care, analogous to the total costs included in the TPCC measure; this broad purpose is conceptually distinct from many other, more narrowly targeted cost and quality performance measures used by CMS. The TPCC measure does not re-estimate these risk adjustment models, but instead uses the risk score coefficients obtained directly from CMS as calculated for use in the Medicare Advantage program.

To test the impact of these social risk factors on the TPCC measure, we estimated analogous regressions using mean monthly cost for beneficiaries included in the TPCC measure as the dependent variable. We analyzed the effects of removing gender and dual from the existing model and adding additional indicators for social risk factors using the AHRQ SES index. We examined the impact of social risk factors in our risk adjustment model by running goodness of fit tests when different risk factors are added or subtracted and compared results to the base risk adjustment model, where the base risk adjustment model refers to the standard set of risk adjustment variables in the CMS-HCC V22 and CMS-ESRD V21 models.

First, we analyzed the model coefficients and p-values for each of the models with and without social risk factors to understand whether any of the social risk factor covariates are predictive of a beneficiary's monthly cost. The T-tests revealed significant p-values for the majority of factors that interact with gender or dual, indicating that social risk factors are likely predictive of resource use among beneficiaries for the relevant characteristic. For the community, institutional, and dialysis models, dual enrollment is associated with systematically higher cost. The addition of AHRQ SES index was significant and negative in value for the community, institutional, and new enrollee models, but not found to be significant in either the dialysis or new enrollee dialysis models. The adjusted R-squared of the models with gender and dual removed decreased by less than 0.005 for the Community, Institutional, New Enrollee, and Community Dialysis models when compared to the base model including these factors. The adjusted R-squared for the New Enrollee Dialysis model decreased by 0.018. When including the AHRQ SES index, there was nearly no difference in adjusted R-squared for all five models. Appendix Tables 2b3.4b.a and 2b3.4b.b present these results.

Secondly, we analyzed the impact of social risk factors on the overall model performance by looking at differences in measure scores calculated with and without social risk factors. Results indicate minor differences in measure score performance, with the removal of gender and dual status having larger effects than the addition of the AHRQ SES index. The measure scores for 87.1 percent of TINs and 85.3 percent of TIN-NPIs changed by ±1 percent or less when gender and dual were removed and compared to the base risk adjustment model. When adding the AHRQ SES index and comparing to the base model, 91.1 percent of TINs and 92.8 percent of TIN-NPIs changed by ±1 percent or less. Scores for nearly 100 percent (>=99.99%) of TINs and TIN-NPIs changed by ±10 percent or less. Appendix Table 2b3.4b.c presents these results in detail.

Finally, we analyzed the correlation between measure scores calculated by risk adjustment models with varied risk factors included. When removing dual and gender and comparing to the base model, the measure scores were highly correlated at both TIN and TIN-NPI levels, with Spearman correlation coefficients of 0.998 for both levels. Similar results were found when adding the AHRQ SES index compared to the base model. These results indicate that the inclusion of additional social risk factors (through the AHRQ SES index) in the current risk adjustment model only has a minor effect on measure scores. Appendix Table 2b3.4b.d presents these results in detail.

Based on the results above, we find no evidence requiring a departure from the standard Medicare Advantage model, which has long incorporated gender and dual status to predict total Medicare costs and which has served as the foundation of the original TPCC measure used in the CMS' Quality and Resource Use Report (QRUR) program for clinicians.

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

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 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.[16,17]

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

The R-squared reported in the December 2018 CMS Report to Congress for the CMS-HCC V22 model for community enrollees, segmented by dual eligibility and disability, range from 0.11 to 0.12. The CMS-ESRD v21 R-squared values are 0.02 and 0.11 for the dialysis new enrollee and dialysis community models, respectively.

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

Included in the standard testing of the HCC and ESRD models in the December 2018 CMS Report to Congress is calibration analyses interpreted as how accurately the risk models' predictions match the actual beneficiary cost. For each of the risk factors included in the models, predictive ratios were calculated as a ratio of predicted cost to actual cost for subgroups of beneficiaries within the model sample to demonstrate the models' prediction accuracy. For all models, the predictive ratio is equal to or close to one for many of the risk factors, indicating that the model is accurately predicting actual beneficiary cost for that risk factor. The detailed results are presented in the series of Table 5-20 of the December 2018 CMS Report to Congress.

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Results of analyses examining predictive ratios by risk decile are included in the December 2018 CMS Report to Congress. Analyses of predictive ratio by risk decile assess the stability of the risk adjustment model among beneficiaries of similar case mixes. As shown in the December 2018 Report to Congress, analyses of these risk deciles for the measure shows that the predictive ratios are generally close to one across all risk score deciles.

For the TPCC beneficiary population, we divided beneficiaries into risk deciles based on their average risk score observed in the performance period and analyzed their mean risk- and specialty-adjusted monthly cost. **Table 9** displays these results, showing the range in mean risk- and specialty-adjusted monthly cost across deciles to be $1,013 to $1,396.

[16] 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.

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

Table 9. Mean Risk- and Specialty-Adjusted Monthly Spending by Risk Decile

Risk Decile (By Mean Beneficiary Risk Score)	Mean Risk-Adjusted, Specialty-Adjusted Cost
1st Decile (Low Risk Score)	$1,205
2nd Decile	$1,138
3rd Decile	$1,173
4th Decile	$1,016
5th Decile	$1,027
6th Decile	$1,013
7th Decile	$1,045
8th Decile	$1,089
9th Decile	$1,195
10th Decile (High Risk Score)	$1,396

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 reported in Sections 3.5.8 and 3.5.9, the predictive ratios for each risk factor included in the model and for all risk deciles are close to one. Predictive ratios close to one indicate that expected spending is accurately predicting observed spending. Additionally, mean risk- and specialty-adjusted cost is within a narrow range across deciles indicating that the model is accurately predicting actual beneficiary resource use and not unfairly penalizing participants based on patient complexity. Overall, the results show that the model is accurately predicting spending, regardless of individual risk factors or overall risk level.

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 TPCC 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 beneficiary-months 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*)

As can be seen in **Figures 1 and 2**, TPCC scores are distributed fairly symmetrically and have a good deal of variability. For the TINs, the standard deviation is $257, and 99/1, 90/10, and 75/25 percentile ratios are 3.49, 1.66, and 1.26, respectively. For the TIN-NPIs, the standard deviation is $310, and 99/1, 90/10, and 75/25 percentile ratios are 3.93, 1.74, and 1.30, respectively.

Figure 1. Distribution of TPCC Scores for TINs

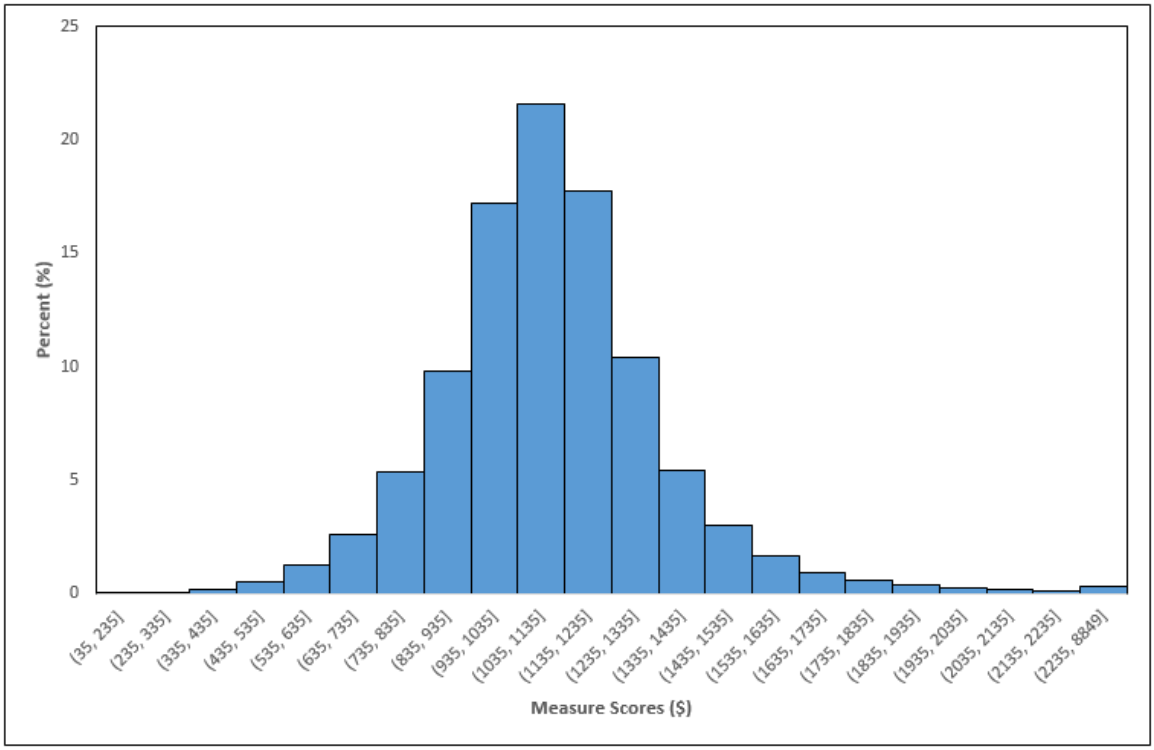
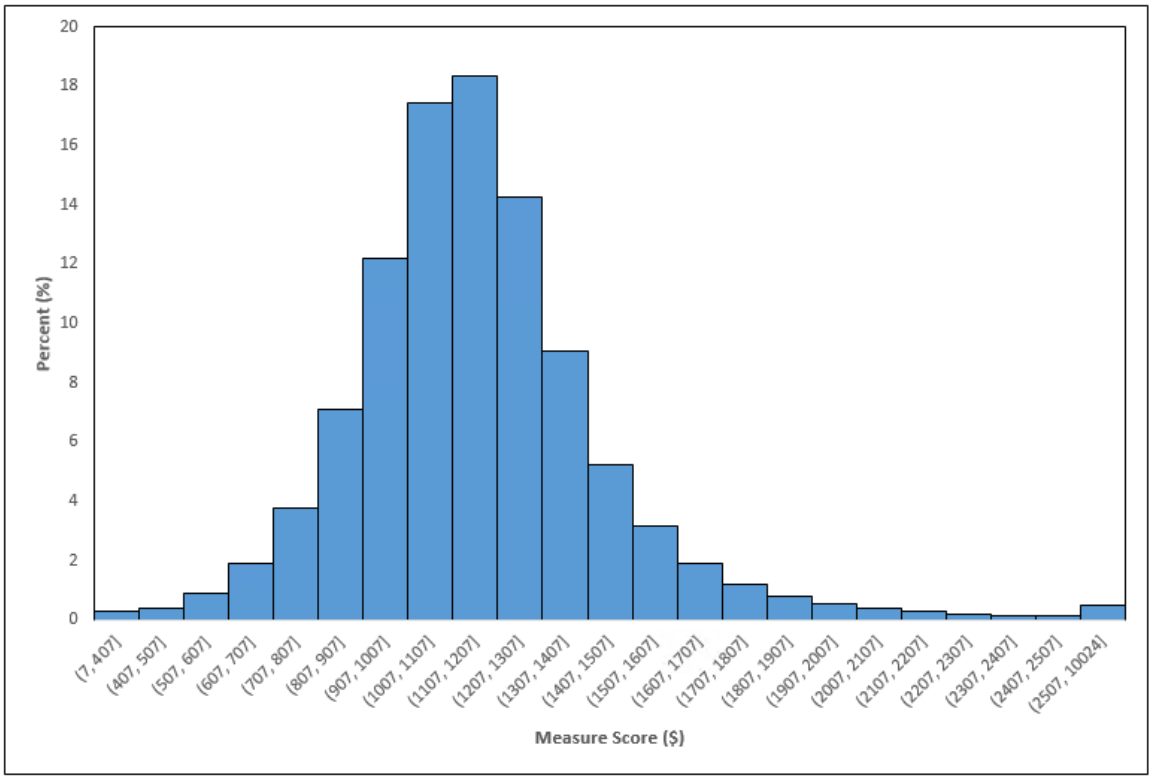


Figure 2. Distribution of TPCC Scores for TIN-NPIs



These results indicate the measure is capturing differences in performance measure scores. The results also show that there are not systemic differences in clinician scores. For instance, the mean scores for clinicians across nine census divisions (excluding 'Unknown') are within a less than $140 range (i.e., $1,058 to $1,197 at the TIN level and $1,136 to $1,233 at the TIN-NPI level). Similarly, clinicians in urban areas seem to perform comparably to those in rural areas on average (i.e., $1,108 in urban compared to $1,117 at the TIN level and $1,170 in urban compared to $1,159 in rural at the TIN-NPI level).

Analysis of clinicians by number of beneficiary months indicates that clinicians with more beneficiaries perform similarly to those responsible for fewer beneficiaries with a max difference in mean score across the various ranges of beneficiary equal to $88 at TIN level and $11 at TIN-NPI levels. 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 $1,074 to $1,204 and a range in mean TIN-NPI score of $1,131 to $1,303, 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 TPCC scores, demonstrated in Section 2a2, small differences in scores can be interpreted as meaningful. This is confirmed by our analysis of statistical significance: 17.9 percent of TINs and 11.4 percent of TIN-NPIs had scores that were statistically significantly higher than the national mean, while 16.8 percent of TINs and 10.9 percent of TIN-NPIs had scores that were statistically significantly lower (**Table 10**).

Table 10. 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	74,191	12,476	16.8%	48,468	65.3%	13,247	17.9%
TIN-NPI	335,480	36,622	10.9%	260,584	77.7%	38,274	11.4%

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 TPCC 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 moderate to high face validity rating from expert clinicians that that scores obtained from the measure specification will provide an accurate reflection of the costs for managing primary care, and can be used to distinguish good and poor performance on cost effectiveness (see Section 2b1.2). For example, empirical results indicate 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.

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 TPCC 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 we excludes beneficiaries enrolled in Medicare Part C or who have a primary payer other than Medicare during the measurement period. In such situations, Medicare Parts A and B claims data may not contain sufficient information to capture the beneficiary's complete clinical risk profile, which is required for 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 11 below presents the frequency of missing data across the three categories of missing data which caused beneficiaries to be excluded from the TPCC measure. Frequency is presented in terms of the number of beneficiaries excluded due to missing data, as well as the number of TINs and TIN-NPIs who had at least one beneficiary excluded due to missing data. The missing data exclusions are:

- Beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the measurement period
- Beneficiary resides outside of the U.S. or Territories
- Beneficiary Enrollment in Medicare for Railroad Workers and their Families

- **Table 11. Missing Data Categories for the TPCC Measure**

Exclusion	# Beneficiaries	# TINs	# TIN-NPIs
No Continuous Enrollment in Medicare Parts A and B, and Any Enrollment in Part C	5,053,189	153,440	802,144
Beneficiary Resides Outside of U.S. or Territories	14,879	17,384	38,890

Beneficiary Enrollment in Medicare for Railroad Workers and their Families	326,588	79,390	392,036

Additional descriptive statistics for exclusions related to missing data are provided in Appendix Table 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 TPCC measure is calculated with Medicare claims data, we expects a high degree of data completeness for those beneficiaries with Medicare Part A and B coverage and 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. If this is an eMeasure, 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 beneficiaries with missing data are excluded to ensure completeness of data and accurate comparability across beneficiary months.

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: Field Testing

Acumen and CMS conducted a national field test of 11 episode-based cost measures and two population-level cost measures, including the Total Per Capita Cost (TPCC), developed during 2018 for a 35-day comment period (October 3, 2018 to November 5, 2018). We provided TPCC Field Test Reports to a sample of eligible clinician groups and clinicians. Each report included information on measure performance for a clinician or clinician group attributed 20 or more beneficiaries. [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 TPCC reports: 567,239
- TIN reports: 120,266
- TIN-NPI reports: 446,973

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 TPCC 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.

Field Testing

During the feedback period, 12,902 field test reports for TPCC were downloaded by 703 clinician groups (TINs) and 12,199 clinicians (TIN-NPIs). Stakeholder comments from field testing were summarized for the TEP to consider in recommending refinements to the measure based on the testing data and feedback.

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

Data Provided During Field Testing

Each TPCC field test report contained the following:

• The clinician or clinician group's measure score along with the national median score and percentile rank

• TPCC cost breakdown by claim type to explain the factors driving the clinician or clinician group measure score (e.g., home health agency, hospice, inpatient, outpatient)

• TPCC cost breakdown by specialty type. The TPCC measure is mostly attributed to primary care physicians and non-physician practitioners, so figures for these two categories are further broken down by specialty (e.g., general practice, family practice, internal medicine, geriatric medicine)

• TPCC 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 a beneficiary-level CSV file that include the risk profile of the attributed beneficiaries.

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 worked with the Physician Value helpdesk and QPP Service Center to answer stakeholder questions during field testing and continued to answer questions after the feedback period ended.

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, "Total Per Capita Cost 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-TPCC.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 we received on measure performance and implementation from the measured entities and others included comments that (i) the revised specifications made several improvements to the current TPCC measure, (ii) while the field test reports and other supplementary materials were helpful, the complexity of these documents was a challenge to some stakeholders, and; (iii) general feedback on the measure's attribution methodology, candidate events, and specialty adjustment. This feedback is detailed in sections U.2.2.2 and U.2.2.3, with references to publicly-available feedback where appropriate.

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 12 comments on the revised TPCC 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 the preliminary recommendation, which received seven comments specific to the TPCC 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 64 results for "tpcc" as a rough approximation of the number of comments on the TPCC 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#COMMENTMUC2018-149MIPS.

U.2.2.2. Summarize the feedback obtained from those being measured.

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:

• Stakeholder engagement and involvement remains an important aspect of the measure development process. Stakeholders expressed appreciation for the opportunity to provide feedback during field testing and for CMS' continued efforts to involve them in the measure development process. Commenters also valued the decision to operationalize previously collected feedback, as demonstrated through the addition of measure-specific workgroups to the development process.

• Field test reports present useful information for understanding clinician performance, though reduced complexity could encourage more clinician participation. Stakeholders praised the presentation and content of the field test reports. However, the complexity of the information presented in the reports was a challenge for some stakeholders.

• Improved supplemental field testing materials are helpful but can be further refined. Some stakeholders found the supplemental field testing materials to be informative and thorough, providing useful information on field testing and the specifications of the cost measures. However, many noted that although the materials are comprehensive, they remain lengthy and complex, and they believe the amount of information provided is too overwhelming to be useful.

• Ample time for review of field testing reports and materials is vital to collecting meaningful stakeholder feedback. Some stakeholders suggested the field testing period be extended or kept open, given the large amount and complexity of the information that was presented.

The report additionally contains measure-specific feedback, which was used as the basis for the post-field testing refinements that were made to the measures, summarized below:

• Refinements to the list of primary care services used as candidate events to ensure they better reflect primary care services.

• Addition of the specialty exclusions so that HCFA specialties who are not identified to be reasonably responsible for providing primary care are not attributed the TPCC measure.

• Ensuring a specialty adjustment is applied to account for costs that vary across specialties and across TINs with varying specialty compositions.

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 revised TPCC measure underwent MAP review during the 2018-2019 cycle. In December 2018, the MAP Clinician Workgroup gave the preliminary recommendation of 'conditional support for rulemaking,' with the condition of NQF endorsement. In January 2019, the MAP Coordinating Committee reversed the Clinician Workgroup's preliminary recommendation and provided a final recommendation of 'do not support for rulemaking with potential for mitigation'. More detail on the mitigating factors is available in the MAP's final report. [1]

[1] "MAP Clinicians 2019 Considerations for Implementing Measures Final Report," National Quality Forum, http://www.qualityforum.org/Publications/2019/03/MAP_Clinicians_2019_Considerations_for_Implementing_Measures_Final_Report.aspx.

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

Careful consideration was given to all feedback gathered during field testing, and several updates were made to the measure based on the recommendations of field testing commenters and TEP comprised of subject matter and measure-development experts.

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 TEP, 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.

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

• Candidate events: Primary care services list was refined to better reflect primary care services, and went from around 5200 codes to 3200 codes. The categories for primary care services have not changed.

• Attributable Clinicians: Excluded clinician from attribution based on their HCFA specialties:

- HCFA specialties eligible for attribution are those that can be reasonably be responsible for providing primary care:
  - Primary care specialties
  - Internal medicine sub-specialties that frequently manage chronic patients with significant conditions in their areas of specialties along with other medical comorbidities
  - Non-physician clinicians who often provide primary care services
- HCFA specialties excluded from attribution were identified as not providing chronic care for significant medical conditions and fall into the following broad categories:

- Surgical sub-specialties
- Non-physicians without chronic management of significant medical conditions
- Internal medicine sub-specialties with additional highly procedural subspecialization
- Internal medicine that practice primarily inpatient without chronic management
- Pediatricians who do not typically practice adult medicine

• Specialty Adjustment: Will be applied based on clinician specialty.

Rulemaking/Public Comment

While the measure did not receive MAP support due to their concerns regarding the revised specifications, CMS believes that the revised measure provides a more appropriate and valid attribution approach than the current TPCC measure used in MIPS and has adequately addressed the mitigating factors outlined by the MAP. For example, CMS has engaged in a range of education and outreach to increase familiarity with the revisions to the measure, including through field testing and national webinars both during and after field testing. The measure has also been tested, including examining how the measure performs at small numbers, and has been found reliable for TINs at various sizes. Testing results are also publicly posted on the MACRA Feedback Page. [1] After consideration of the public comments, the revised TPCC measure was finalized as proposed.

[1] 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.

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)

1604 : Total Cost of Care Population-based PMPM Index

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?
No

H.2.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.
TCOC is tested and endorsed for a population of patients less than 65 years of age, while TPCC was developed and tested on the Medicare population, affecting the appropriate intended use of each respective measure.

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.)
N/A.
There are no competing NQF-endorsed or non-NQF-endorsed cost measures that address the same measure focus and target population.

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

Jackson 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

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: