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 #:** 3564

**De.2. Measure Title:** Medicare Spending Per Beneficiary – Post Acute Care Measure for Home Health Agencies

**Co.1.1. Measure Steward:** Centers for Medicare and Medicaid Services

**De.3. Brief Description of Measure:** The Medicare Spending Per Beneficiary – Post Acute Care Measure for Home Health Agencies (MSPB-PAC HH) was developed to address the resource use domain of the Improving Medicare Post-Acute Care Transformation Act of 2014 (IMPACT Act). This resource use measure is intended to evaluate each home health (HH) agency's efficiency relative to that of the national median home health agency (HHA). Specifically, the measure assesses Medicare spending by the HHA and other healthcare providers during an MSPB-PAC HH episode. The measure reports the ratio of the payment-standardized, risk-adjusted MSPB-PAC Amount for each HHA divided by the episode-weighted median MSPB-PAC Amount across all HHAs. The MSPB-PAC Amount is the ratio of the observed episode spending to the expected episode spending, multiplied by the national average episode spending for all HHAs. The measure is calculated using two consecutive years of Medicare Fee-for-Service (FFS) claims data and was developed using calendar year (CY) 2015-2016 data. This submission is based on CY 2016-2017 data; i.e., HHA admissions from January 1, 2016 through December 31, 2017.

Claims-based MSPB-PAC measures were developed in parallel for the HH, inpatient rehabilitation facility (IRF), long-term care hospital (LTCH), and skilled nursing facility (SNF) settings to meet the mandate of the IMPACT Act. To align with the goals of standardized assessment across all settings in PAC, these measures were conceptualized uniformly across the four settings in terms of the construction logic, the approach to risk adjustment, and measure calculation. Clinically meaningful case-mix considerations were evaluated at the level of each setting. For example, clinicians with HH experience evaluated HH claims and then gave direction on how to adjust for specific patient and case-mix characteristics.

The MSPB-PAC HH measure was adopted by the Centers for Medicare & Medicaid Services (CMS) for the HHA Quality Reporting Program (QRP) and finalized in the CY 2017 Home Health Prospective Payment System Rate Update; Home Health Value-Based Purchasing Model; and Home Health Quality Reporting Requirements.[1]

Public reporting for the measure began in Fall 2018 through the Home Health Compare website (https://www.medicare.gov/homehealthcompare/search.html) using CY 2017 data.

Notes:

[1] Medicare and Medicaid Programs; CY 2017 Home Health Prospective Payment System Rate Update; Home Health Value-Based Purchasing Model; and Home Health Quality Reporting Requirements. Federal Register, Vol. 81, No. 213. https://www.govinfo.gov/content/pkg/FR-2016-11-03/pdf/2016-26290.pdf

**IM.1.1. Developer Rationale:** MSPB-PAC HH QM was developed to address the resource use domain of the Improving Medicare Post-Acute Care Transformation Act of 2014 (IMPACT Act). As part of the IMPACT Act, MSPB-PAC aims to achieve interoperability, data exchange, and standardized measurement among post-acute providers. The mandated use of MSPB-PAC measures is intended to allow for a greater ability to measure resource use and efficiency of care to improve outcomes, as well as encourage all PAC providers towards aligned incentives and care coordination.

Differences in post-acute care payments are a key driver of variation in Medicare spending overall.[1,2] In addition, there have been a number of studies demonstrating significant variability in HH care and outcomes, links between provider characteristics and readmissions, and significant opportunities for improvement.[3,4,5] The cost and quality link is important, with this resource use measure playing an important role in discerning value of HH care.

The MSPB-PAC HH measure was adopted by CMS for the HH Quality Reporting Program (QRP) and finalized in the CY 2017 HH Prospective Payment System (PPS) Final Rule.[6] Public reporting for the measure began in Fall 2018 through the HH Compare website.

[1] Institute of Medicine. (2013). Variation in Health Care Spending Assessing Geographic Variation. (July)

[2] Kahn, E. N., Ellimoottil, C., Dupree, J. M., Park, P., & Ryan, A. M. (2018). Variation in payments for spine surgery episodes of care: Implications for episode-based bundled payment. Journal of Neurosurgery: Spine, 29(2), 214–219.

[3] Murtaugh C.M., Deb P., Zhu C., Peng T.R., Barrón Y., Shah S., Moore S.M., Bowles K.H., Kalman J., Feldman P.H., Siu A.L. (2017). Reducing Readmissions among Heart Failure Patients Discharged to Home Health Care: Effectiveness of Early and Intensive Nursing Services and Early Physician Follow-Up. Health Services Research. 52(4), 1445-1472.

[4] Lohman M.C., Cotton, B.P., Zagaria, A.B., Bao, Y., Greenberg, R.L., Fortuna, K.L., Bruce, M.L. ( 2017). Hospitalization Risk and Potentially Inappropriate Medications among Medicare Home Health Nursing Patients, Journal of General Internal Medicine. 32(12), 1301-1308.

[5] Institute of Medicine. (2013). Variation in Health Care Spending Assessing Geographic Variation. (July)

[6] Medicare Program; Calendar Year 2017 Home Health Prospective Payment System Update; Home Value-Based Purchasing Model; and Home Health Quality Reporting Requirements for Federal Register, Vol. 81, No. 213. https://www.govinfo.gov/content/pkg/FR-2016-11-03/pdf/2016-26290.pdf

**De.1. Measure Type:** Cost/Resource Use

**S.5. Data Source:** Assessment Data

Claims

Enrollment Data

Other

**S.3. Level of Analysis:** Facility

## 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 intended to evaluate each home health (HH) agency's efficiency relative to that of the national median home health agency (HHA). The measure assesses Medicare spending by the HHA and other healthcare providers during an MSPB-PAC HH episode and reports the ratio of the payment-standardized, risk-adjusted MSPB-PAC Amount for each HHA divided by the episode-weighted median MSPB-PAC Amount across all HHAs.
- It was developed to address the resource use aspect of the Improving Medicare Post-Acute Care Transformation Act of 2014 (IMPACT Act) to allow for a better ability to measure resource use and efficiency of care to improve outcomes and align incentives and care coordination across PAC providers.
- This measure is calculated using two consecutive years of Medicare Fee-for-Service (FFS) claims data and was developed using calendar year (CY) 2015-2016 data.
- The MSPB-PAC Amount is the ratio of the observed episode spending to the expected episode spending, multiplied by the national average episode spending for all HHAs.
- The developer showed differences in in post-acute care payments are a key driver of variation in Medicare spending overall, showing studies demonstrating significant variability in HH care and outcomes, links between provider characteristics and readmissions, and significant opportunities for improvement.

**1b. Opportunity for Improvement.**

- The developer showed MSPB-PAC HH measure scores reported publicly for all US providers paid under Medicare's HH Prospective Payment System (PPS) with 20 or more eligible episodes in the reporting period. There were a total of 10,470 HHAs with 20 or more episodes in CY 2016-2017. These scores represent 10,321,802 patient episodes, after all exclusions were applied. The scores variability included a mean of 0.96, standard deviation: 0.15, with a minimum of 0.31 and maximum of 2.44 and an interquartile range of 0.18.

**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:**

- **The measure is part of a Congressionally mandated set of measures on post-acute care. Legislative requirement probably creates adequate impact**
- **Yes**
- **Yes.**
- **yes, there is variation in spending and it affects a sizeable fraction of the Medicare population**
- **The developer describes overall variation in cost in post-acute care. I agree this is an area for improvement. However, the cost measures presented are for each of the different post-acute care settings. A more effective approach may be to allow providers to identify lower cost post-acute care settings to send a patient rather than trying to control costs within a specific setting.**
- **Yes (required by IMPACT Act)**
- **No concerns**
- **Yes, high resource use/cost**

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

**Comments:**

- **Variation in observed to expected ratios sufficiently large to suggest opportunity for improvement. .35 difference over 10th-90th percentile.**
- **Yes**
- **Using data from calendar year 2015-2016, the developer estimated mean MSPB-PAC HH measure score of 0.96 (SD=0.15, Min = 0.31, Max = 2.44, and IQR = 0.18) for all US providers paid under Medicare HH PPS with ≥20 eligible episodes during the reporting period. The range of scores indicates variations in costs which can be improved.**
- **yes, there is some room.**
- **Although the measure has been publicly reported since Fall 2018, since this is the measure's initial endorsement, no improvement data was provided. It is unclear whether public reporting has impacted HH costs.**

- **There were a total of 10,470 HHAs with 20 or more episodes in CY 2016-2017. These scores represent 10,321,802 patient episodes, after all exclusions were applied. The scores variability included a mean of 0.96, standard deviation: 0.15, with a minimum of 0.31 and maximum of 2.44 and an interquartile range of 0.18.**
- **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-3; M-3; L-1; I-1
Methods Panel Individual Validity Ratings: H-3; M-3; L-1; I-1

- The developer provided responses to the concerns raised by the SMP, which can be found in the SMP Spring 2020 Discussion Guide on page 87 – 89.

**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 home health episodes.

- The performance measure score reliability testing was based on 10,470 home health agencies (HHAs) in the measurement period of 2016-2017.
- The developer reported that the mean reliability score for all agencies was 0.84 with median of 0.90. When examined by facility size, the average reliability score ranged from 0.63 (Q1) to 0.97 (Q4). The ICC for the overall sample was 0.76 with 95% confidence interval of 0.75-0.77. The ICC was lowest for Q1 (0.57) and highest in Q4 (0.94).

**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:**

- **This measure excludes costs for "Clinically Unrelated Services." The process of aggregating costs to be considered for exclusion and the expert panel process are relatively clear. In the process, services that did not account for a sufficiently large share of payments within their respective clinical service category were not included in the review to allow clinicians to focus their review on services representing a higher percentage of overall Medicare spending within the episode window. The basis for determining which costs were not examined under this criterion is not specified. I am concerned that by looking at aggregates, costs that substantially and idiosyncratically affect individual patients may not be considered, and low volume providers will be penalized by random events unrelated to their care. I would like more description of this process than is provided in Appendix D. THIS IS A COMMON ISSUE ACROSS THE MEASURES BEING CONSIDERED IN THIS CYCLE.**
- **Not clear how much discretion HH providers have over the healthcare utilization included, so it is not clear whether these measures truly related to HH provider performance. Potential of and handling overlapping episodes for each patient is unclear. How is death handled?**
- **The comment from Panel Member #3 of the Scientific Methods Panel about the use of "other data" from the Minimum Data Set (MDS) needs further discussion. Furthermore, the panel members have other important questions/concerns the definitional transparency of this measure (e.g., attribution of services to specific episode in case of multiple episode during measurement period, winsorization of the outliers). These need to be discussed at the committee discussion.**
- **missing data not a problem. data elements are defined. complicated to construct.**
- **No concerns that were not raised by the Scientific Methods Panel.**
- **none**
- **No concerns**
- **No concerns**

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

**Comments:**

- **S/N ratios in first quartile are too low for acceptable reliability: Reliability unacceptable in 1st quartile: 20-180 (SN: 0.52-0.76, ICC: 0.57) and at bottom end we have previously found acceptable for 2nd quartile (SN: 0.81-0.89, ICC: 0.82)**
- **None**
- **While the reliability assessment methods are valid, reliability varies across the provider size, potentially making it difficult to differentiate home health agencies with smaller number of qualifying episodes (panel member # 2 & # 3). Furthermore, the comment from panel member #3 is noted about potentially incorrect data source, and needs committee-level discussion.**
- **the split sample showed reproducablity. the reliablity testing showed that there was sufficient to high reliability to differentiate provider performance (at 0.7 or higher)**
- **No concerns that were not raised by the Scientific Methods Panel.**
- **none**
- **No concerns**
- **No concerns**

## Validity

**2b1. Specifications align with measure intent:**
The measure specifications are consistent with the measure intent and captures the most inclusive target population.
**2b2. Validity Testing:**
Demonstration that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided.
**2b3. Exclusions:**
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. Risk Adjustment:**
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. <u>Disparities</u>:** 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 HHAs. This attribution approach was developed in order to drive HHAs to facilitate care coordination and support their role in 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 three separate empirical tests on validity:
    - 1) Assessed how this measure correlates to resource /utilization such as hospitalization within the episode window and emergency room (ER) visits within the episode window.
    - 2) Correlated this measure with the Discharge to Community (DTC) rates for HHAs (measure endorsed by NQF (#3477).
    - 3) Correlated this measure with the Acute Care Hospitalization (ACH) During the First 60 Days of Home Health are endorsed by NQF (#0171).
    - 4) Correlated this measure with provider's functional improvement quality measure scores publicly reported on the Home Health Compare website for CY 2017. Specifically, the developers used the following NQF-endorsed assessment-based measures:
        - Improvement in ambulation (#0167)
        - Improvement in bathing (#0174)
        - Improvement in bed transfer (#0175)
        - Improvement in management of oral medications (#0176)
        - Improvement in pain interfering with activity (#0177).
- The developers found a positive relationship between MSPB and known indicators of resource or service utilization.
    - The mean observed to expected cost ratio for episodes without a hospital admission is 0.68, compared with 2.31 for episodes with at least one hospital admission during the episode period (p-value<0.0001).
    - The mean observed to expected cost ratio for episodes without an ER visit is 0.89, compared to 1.39 for episodes with at least one ER visits (p-value<0.0001). They also observed a positive relationship between the mean observed to expected cost ratio and the number of hospitalizations/ER visits as hypothesized.
- The developers found a small but significant negative association between MSPB measure scores and the DTC measure scores as hypothesized and a very small but statistically significant correlation (Pearson -0.240; Spearman -0.250) between MSPB measure scores and DTC measure scores
- The developers found a small positive correlation between MSPB measure scores and ACH scores (Pearson 0.298; Spearman 0.305).
- Lastly, the developer found a small, but significant positive correlations between MSPB scores and the various functional improvement scores as hypothesized (Pearson correlations ranging from 0.075 to 0.163; Spearman ranged from 0.041 to 0.152).

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

- The developer reports 19.8% of episodes were excluded because of one or more exclusion criteria

**2b4/2c. Risk adjustment**

- The developer uses a linear regression risk adjustment model with over 124 risk factors.
- The developer reported results showing that spending was slightly higher for patients who were dual eligible ($11,926 vs $11,621 for non-duals), non-white ($12,060 for Blacks compared to $11,739 for whites), and had low socioeconomic status (SES) based on AHRQ SES index ($11,962 in Q2 of SES Index vs $11,553 in highest quantile).
- The developer reported that each of the social risk factors was statistically significant in the risk adjustment model. However, the developer did not include them in the overall model, concluding that adding them, individually or together, did not substantially improve overall model fit.
- The developer stated that the social risk factors had minimal impact on average measure scores, so they chose to not adjust the scores for these social risk factors.
- The overall R-squared was 0.092. The observed to predicted costs varied between 0.96 and 1.05 for the ten deciles.

**2b5: Meaningful Differences**

- Across 10,470 facilities, the mean score is 0.96 with a standard deviation of 0.15, and a minimum score of 0.31- with a maximum of 2.44, with. The 10th- percentile was 0.78 and the 90th percentile distribution was 1.13.

**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 agree with the developer's rationale for not included SES factors in the risk-adjustment model?*

**Guidance from validity algorithm:**

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

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

**Committee Pre-evaluation Comments:**
**Criteria 2b: Validity**

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

**Comments:**

- **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.**
- **Not clear if result is indicitive of HH performance or the local area healthcare system.**
- **No concerns.**
- **Concern about including measures that are part of MSPB as part of validity testing. Does show somewhat stronger correlation (though still low) with the quality measures, but the findings were statistically significant. R2 is lower than the other MSPB measures (0.09), so much of variation not being explained**
- **I am concerned the R-squared for this model is very small. In addition, the magnitude of difference in payments is also small albeit statistically significant. This could be a function of the large number of episodes being able to detect even small differences in cost. In terms of face validity, I also question whether home health would be able to prevent or contribute to hospital admissions and ED use. The developers do not provide data on the ability for home health to manage the patient's care or whether their care is what led to the admission or ED visit.**
- **none**
- **No concerns**
- **No concerns**

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

**Comments:**

- **None**
- **Whose performance does the measure describe? How is death handled?**
- **None.**
- **none**
- **No concerns not raised by the Scientific Methods Panel.**

- **none**
- **No concerns**
- **No concerns**

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

**Comments:**

- **okay**
- **None.**
- **missing data not a problem**
- **None.**
- **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:**

- **Empirical validation was consistent with past reviews but limited and incomplete. I have two main problems with this measure: First, while the concept is clear (all Part A and B Medicare expenses during the SNF visit and 30 day window after), there is no extended discussion of services actually included, and the major sources of variation in expenses across episodes. Some are specifically discussed, such as hospital readmission and ED, but I would like to better understand what differences in spending are systematically associated with higher or lower costs across patients and facilities to better assess whether these are under the control of the HHA or influenced by the care the HHA provides. I would also like to see more empirical testing of other sources of variation. I am concerned there are real issues of face validity of this measure. The underlying issue for these measures is whether the care decisions made by the provider, in this case the HHA, contribute to the costs of care. There are two pathways for this: "profligate spending," i.e., high spending on a variety of services with low value for patients that run up cost with little gains in patient health or comfort, and complications to which the provider contributes that require treatment and add to cost. There is no discussion of the role of HHAs in preventing or contributing either to hospital admissions or ED use or any other high cost items that run up the expense during the period of treatment and 60 days after treatment ends for Standard Episodes and LUPA episodes. There is a certain element of cookie cutter measure development reflected in this measure -- 60 days after, standard risk adjuster—rather than tailoring to the service so that real cost variances associated with how the service is delivered are identified.**
- **None**
- **Episodes with residuals below 0.1 percentile or above 0.99 percentile of the residual distribution are excluded. Scientific Methods Panel members #3 and #9's observations on handling of outliers by this measure need further discussion. Specifically, should the measure necessarily exclude outliers at all? The R-squared of 0.1 (rounded) seems low for the risk adjustment models with such large samples and so many adjustors. It seems to suggest that models may not have incorporated some key predictors of costs because they are potentially not observed in the data.**
- **outliers shouldn't be dropped but winsorized**
- **It is unclear whether the accountable entity is providing appropriate care although potentially tied to higher costs.**
- **none**
- **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:**

- **Part C exclusion inherently leads to large proportion of excluded cases.**
- **Agree with the concerns expressed by Panel Members #4 and #8 on the overlapping episodes and their correct attribution to multiple providers.**
- **none**

- **There is no comparison to patients that may have been treated through home health but instead were treated in a different post-acute care setting.**
- **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:**

- **Risk adjustment approach is standard approach for CMS measures. It appears to take into account underlying differences in disease/chronic conditions/age. Social risk factors analyzed, and as implemented in risk adjustment model, make small-negligible difference in score or adjusted R-square. I have concerns about use of zip code level SES index, rather than census tract/address based adjustment, and we have no data on how much variability across facilities there is in social risk factors, and whether facilities with high proportion of disadvantaged SES patients regularly score worse, and whether the differences in scores are larger for these facilities. The risk adjuster explains a small proportion of variance in cost. This is not unusual with cost measures. But the risk adjustment model is pretty much a cookie cutter of the standard risk adjustment model for CMS cost measures, and I don't see any thought given to what is unique about patients in HH, and what are the drivers of differences in costs that are not under the control of the HHA providers, to assure these are controlled for. There is also boiler plate language about not wanting to include too many interactions because of risk of overfitting in a multimillion record data set where the potentially interacted measures and cases with both or neither conditions are probably with adequate numbers that overfitting is not a problem. The key line is consistency with the standard CMS approach when we should be asking whether the risk adjustment model is appropriate for controlling for cost and use differences across HHAs that are not under the control of HHA.**
- **As suggested by reviewer, risk-adjustment models should be based on theory, not empirics. Dropping variables will bias the coefficient estimates of the remaining variables**
- **While the measure developer offers rationale for not including SES in the final measure, I agree with the Scientific Panel Member #2 as to why the measure should actually include SES. Exclusion of SES may potentially penalize HHAs that serve higher proportion of patients with these SES factors.**
- **effects of risk adjustment for SES seems very modest in this setting**
- **None.**

- **Developer reported results showing spending was slightly higher for patients who were dual eligible ($11,926 vs $11,621 for non duals for the full population, or 2.6% higher which can be a significant difference in the HH Medicare payment world, non-white (e.g., $12,060 for Blacks compared to $11,739 for whites, or 2.7% higher which can have very significant impact on a plan with many non-white members, and had low SES based on AHRQ SES index ($11,962 in Q2 of SES Index vs $11,553 in highest quantile, or 3.5% higher), and that each of those social risk factors was statistically significant in the risk adjustment model, they concluded that adding them, individually or together, did not substantially improve overall model fit. The SES factors seem to align, yet developer chose not to include them in the overall model.**
- **No concerns**
- **Several methods reviewers have raised concerns about the lack of inclusion of social risk factors and baseline functional status. Newer data has come out with respect to Social Risk Factors and impact on outcomes. This should be reconsidered.**

## Combined Scientific Methods Panel Preliminary Analysis of Scientific Acceptability

**Measure Number:** 3564

**Measure Title:** Medicare Spending Per Beneficiary – Post -Acute Care Measure for Home Health Agencies
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: Minimum Data Set (Panel Member #2)**

**Panel member #3** NOTE: The Developer submitted form indicates that they gather some "Other" data from the Minimum Data Set (MDS). To the best of my knowledge, home health agencies do NOT complete the MDS for their patients. The standardized patient clinical assessment form is the OASIS instrument. Additionally, this error (i.e., referencing the MDS as the source of numerous risk factors used in the prediction model) is repeated throughout the submitted document.

This error may be the result of a "copy/paste" error from the other related measures (i.e., 3561-3563). However, the effect of the error is to make the information submitted not applicable to any measure related to home health agencies. Hence, because the source of the underlying data used to compute the MSPB measure for home health agencies is not properly specified, the quality of the reliability and validity of the MSPB measure for home health agencies cannot be properly determined.

**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** See previous note regarding the source of home health agency data (i.e., OASIS, not MDS).

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

   Exclusions of unrelated services is briefly described but Appendix D, with extensive description not provided to committee

   **Panel Member #1** None

   **Panel Member #2** This is a very complex measure to calculate requiring many steps, some of which seem a bit ambiguous, such as which events apply as exclusions, where episodes end, which events be included in 2 different overlapping episodes, etc. While CMS may be able to utilize the developing set of codes and logic, I do not think this measure is easily replicable by others who may want to calculate and compare HOME HEALTH spending. It likely has applicability only to the CMS HOME HEALTH compare program for public reporting.

   **Panel Member #**3 SEE RELATED COMMENTS REGARDING TERMS, MEASURE SCORE, MEASURE IN CURRENT USE, APPROPRIATE USE OF DATA, TIME PERIOD FOR PERFORMANCE, STANDARDIZING COST OF CARE, AND CLINCIAL DIFFERENCES AMONG PATIENTS ACROSS FACILITIES FROM MY REVIEW OF MEASURE 3561.

   **Panel Member #7** No concerns

   **Panel Member #8** General concern about the potential overlap in the various MSPB metrics submitted. It appears that each services may be attributed to multiple 'episodes' and providers. Unclear how improvements may be made when the attribution is so scattered.

   **Panel Member #9** My key concern is with the outlier exclusion (Step 6 – Contruction logic). Based on the measure specifications, episodes with residuals below the 1$^{st}$ percentile or above the 99$^{th}$ percentile of the residual distribution are excluded. Four factors make this approach particularly concerning: 1) This is a measure focusing on resource utilization, should episodes with very low or very high residuals be excluded? If the concern is with potentially undue influences of outliers, is exclusion the best available approach? 2) Winsorize predicted values: low predicted value below 0.5$^{th}$ percentile is already winsorized before calculation of residual. 3) Closing episodes: full payment for all claims that begin within the epidose window is counted toward the episode. This may make it likely that such episodes (with substantial claims at the end of episode window) become outleirs and be excluded. 4) There are substantial differneces between three different episode types, outlier exclusion based on residuals may favor one type of episode. At minimum, the developer should report the distribution of outlier exclusion across facilities to ensure that they don't concentrate in a limited number of facilities.

### 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 VALIDITY testing** of patient-level data conducted?

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

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

   **Panel Member #3** 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 (reliability score) and split-sample reliability testing (ICC).

   Split sample reliability

   Differences tested across 4 quartiles of size, with most critical the bottom 2 quartiles (20-180, 181-466)

   **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 facility based on randomly-spllit, independent subsets of HOME HEALTH episodes. Good agreement indicates the performance score is more the result of facility characteristics (efficient care) than statistical noise due to random variation. They used 4 years of data to achieve #'s of episodes per facility comparable to the numbers used for actual measurement (at least 20 episodes per year) with episodes across years evenly distributed. They used the Shrout-Fleixx interclass correlation coefficient (ICC) between the split-half scores to measure reliability.

   **Panel Member #3** Methodology (Reliability Score and split-sample Shrout-Fleiss ICC) is appropriate.

   **Panel Member #7** Split sample reliability testing was utilized as well as intraclass coefficient. The quartile means and ranges increased from Quartile 1 to Quartile 4.

   **Panel Member #8** ICC using split half.

   **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** Based on high reliability score (mean score = 0.84) and ICC of 0.76 overall, and by sample size quartile, the measure exhibits high reliability.

**Panel Member #2** Overall, reliability testing using 10,470 HHAs from 2016-2017 indicated good reliability., regardless of facility size. The average reliability score for all agencies was 0.84 with median of 0.90. When examined by facility size, the average reliability score ranged from 0.63 (Q1) to 0.97 (Q4). The ICC for the overall sample was 0.76 with 95% confidence interval of 0.75-0.77. The ICC was lowest for Q1 (0.57) and highest in Q4 (0.94)

**Panel Member #3** In addition to the cavaet regarding data source noted previously, the Reliability Score values show a distinct difference based on HHA size as shown in the following table:

**Table 3. Agency Reliability Score Distribution of the Episode-Level MSPB Risk Adjusted Spending, overall, HHA sample and by sample size quartile, with public reporting exclusions (k = 10,470)**

| Agency Sample | K | Mean (SD) | 25th Pct* | Median | 75th Pct |
|---|---|---|---|---|---|
| Overall | 10,470 | 0.84 ( 0.16) | 0.79 | 0.90 | 0.95 |
| Quartile 1: 20-180 episodes | 2,623 | 0.63 ( 0.18) | 0.52 | 0.66 | 0.76 |
| Quartile 2: 181-466 episodes | 2,619 | 0.84 ( 0.06) | 0.81 | 0.85 | 0.89 |
| Quartile 3: 467-1,063 episodes | 2,612 | 0.92 ( 0.03) | 0.91 | 0.92 | 0.94 |
| Quartile 4: 1,064-68,523 episodes | 2,616 | 0.97 (0.02) | 0.96 | 0.97 | 0.98 |

Reliability unacceptable in 1st quartile: 20-180 (SN: 0.52-0.76, ICC: 0.57) and at bottom end we have previously found acceptable for 2nd quartile (SN: 0.81-0.89, ICC: 0.82)

* Pct = percentile. Analysis of Medicare Claims File for HH CY 2016-2017.

A similar pattern is found in the split-sample ICC results:

**Table 4. Split-sample reliability: Intraclass correlation coefficients between split-sample performance measure scores for the overall HHA sample and by sample size quartile, with public reporting exclusions (N = 10,470)**

| Agency Sample | K | ICC(2,1) (95% CI) |
|---|---|---|
| Overall | 10,470 | 0.76 (0.75-0.77) |
| Quartile 1: 20-180 episodes | 2,623 | 0.57 (0.54-0.59) |
| Quartile 2: 181-466 episodes | 2,619 | 0.82 (0.81-0.83) |
| Quartile 3: 467-1,062 episodes | 2,610 | 0.90 (0.89-0.90) |
| Quartile 4: 1,063-68,523 episodes | 2,618 | 0.94 (0.94-0.95) |

Analysis of Medicare Claims File for HH CY 2014-2017.

**Panel Member #7** Facility reliability score had a mean of 0.84 overall with a range of 0.52 for the 25th percentile and 0.96 for the 75th percentile. Split sample had an intraclass coefficient of 0.76 overall, with a range of 0.57 for the first quartile to 0.94 for the fourth quartile.

**Panel Member # 9** Results are very 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** (methodology was acceptable; the results showed problems)☐

☐ **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** The rationale has been stated in 6 and 7 above.

    **Panel Member #2** Reliability testing using two different approaches shows moderate to high reliability, but reliability was considerably lower for smaller agencies (0.63 reliability score and 0.57 ICC for agencies with 20-180 epidoses. This higher variability of scores based on sample could result in less ability to differentiate good providers with smaller number of episodes.

    **Panel Member #3** Given the data source problem identified previously, I rated the reliability as "insufficient." If the data source problem is addressed satisfactorily, then the reliability rating would be changed to "low" given the wide variation in reliability between smaller and larger HHAs.

    **Panel Member #4** This is another measure that uses Medicare claims data. While these data are audited to assure billing is accurate, we don't have independent assessment of data element quality.

    Reliability in 1st quartile not acceptable. At minimum, need to use a higher threshold for number of cases.

    **Panel Member #7** Quartiles show significant variation in the reliability scores and in intraclass coefficients but remained moderate to good.

    **Panel Member #8** NA-

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

## VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    **Panel Member #1** None

    **Submission document:** Testing attachment, section 2b2.

**Panel Member #2** NONE

**Panel Member #3** Fewer than 20% of home health episodes were eliminated due to Medicare payment exclusions. The performance is better than in many other post-acute care settings.

**Panel Member #4** Case exclusion criteria are reasonable, although 11.6% excluded for "Any episode in which a beneficiary is not enrolled in Medicare FFS for the entirety of the 90-day lookback period (i.e., a 90-day period prior to the episode trigger) plus episode window (including where a beneficiary dies), or is enrolled in Part C for any part of the lookback period plus episode window" Would be useful to know how many of these are Medicare<90days, part C, or deaths.

**Panel Member #7** No concerns

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

**Panel Member #1** None

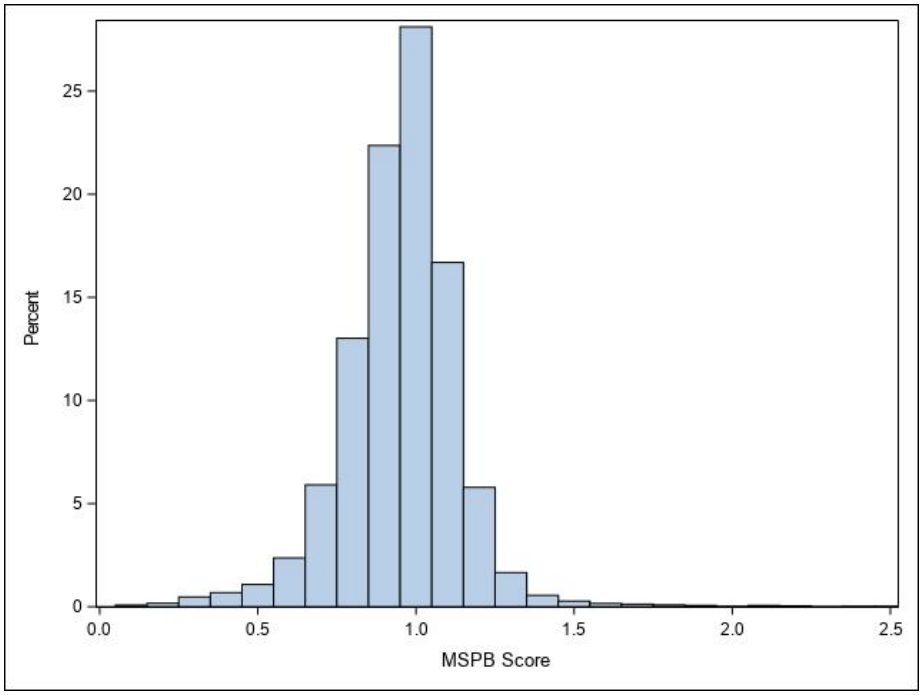
**Submission document:** Testing attachment, section 2b4.

Panel Member #2 NONE

**Panel Member #3** Used a bootstrapping methodology to generate samples of HHAs. Why? With nearly 10K HHAs and more than 10 million episodes of care, there should be sufficient data available to determine if there are meaningful differences among these HHAs with creating an artificial set of data.

Divided HHA MSPB performance into three groups: Below national MSPB (34.6%), At national MSPB (23.4%), and Above national MSPB (42.0%) based on an unspecified "statistical significance" value. The presented distribution looks relatively normal:

**Figure 3. Distribution of MSPB-PAC HH Scores**




Analysis of Medicare Claims File for HH CY 2016-2017.

Why is the three-group stratification U-shaped? If the "statistical significance" values was based on 1 standard deviation above or below the national mean, then the proportions would be dramatically

different based on the presented distribution (approximately 15/70/15 vs. 35/23/42). Why this dramatic difference? Is this how HHAs are being compared based on the MSPB score?

**Panel Member #4** None. The distribution of risk adjusted predicted expenses are narrower than actual, and relatively tight, so measure generates meaningful differences. Ratios of actual to expected range from 0.31-2.44, with 10th-90th percentile distribution 0.78-1.13.

**Panel Member #7** Across 10,470 facilities, the mean score is 0.96 with a standard deviation of 0.15, and a mimimum score of 0.31- with a maximum of 2.44, with. The 10th- percentile was 0.78 and the 90th percentile distribution 0.78-1.13.

**Panel Member #9** No concern.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
**Panel Member #1** N/A
**Submission document:** Testing attachment, section 2b5.

**Panel Member #1** NA
**Panel Member #2** NONE
**Panel Member #3** No comment
**Panel Member #9** No concern.

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

**Panel Member #2** None

**Submission document:** Testing attachment, section 2b6.

**Panel Member #3** NA. Only small portion of data missing.

**Panel Member #2** NONE

**Panel Member #3** No comment

**Panel Member #7** Missing (problematic) data was present in 0.21% of the episodes and were excluded

**Panel Member #9** No concern.

16. **Risk Adjustment**

16a. **Risk-adjustment method ☐ None ☐☒ Statistical model ☒(124 variables) RFs) ☒ ☒☐ Stratification Panel Member #3 (3 risk categories—HH Standard, Low Utilization Payment Adjustment (LUPA), and Partial Episode Payment (PEP))**

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

**Panel Member #3** Developers' analyses show that there was a relationship between socio-demographic risk factors and MSPB measure:

- Overall, 31% of HHA episodes in the CY 2016-2017 period involved patients who were dual-eligible,

- LUPA and PEP episodes are somewhat more likely to involve patients who are not dual-eligible, are White, and reside in areas with higher average SES.
- PEP episodes are also more likely to involve patients living in urban areas.
- we found that spending is higher for patients who are dual-eligible, patients who are Black, and patients who live in urban areas.
- Spending is highest among patients who reside in areas with average SES in the second and third quartiles of the AHRQ SES Index
- We found that each SRF is statistically significant, when added to the model individually as well as when added together with all other SRFs

Despite these findings, the Developers chose to NOT include socio-demographic risk factors: Given the minimal impact of including SRFs in the risk adjustment model on measure scores, we do not recommend adjusting the scores for these social risk factors.

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 E.g., hospice X—not applicable **(Panel Member #3)**

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

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

**Panel Member #3** The overall adjusted R-Square values for predicting MSPB measure values was poor:

The overall adjusted R-squared is 0.092. The overall adjusted R-squared was computed by fully interacting episode type (LUPA, PEP, or standard) with all other coefficients. This shows the combined explanatory power of the strata (episode types) and the covariates. The adjusted R-squared for the HH Standard episodes is 0.090; it is 0.096 for HH LUPA episodes and 0.076 for HH PEP episodes.

The episode-level predicted costs range from $2,562 to $69,577.

**Table 11. HH Model Diagnostics: Comparison of Observed and Predicted Spending by Predicted Spending Deciles**

| Deciles of predicted episode cost | Number of episodes | Observed episode cost | Predicted episode cost | Predicted minus observed cost | Observed/predicted costs |
|---|---|---|---|---|---|
| 1 | 1,045,627 | 6632.12 | 6328.02 | -304.10 | 1.05 |
| 2 | 1,021,016 | 7394.15 | 7260.85 | -133.30 | 1.02 |
| 3 | 1,027,795 | 7939.56 | 7904.32 | -35.24 | 1.00 |
| 4 | 1,031,299 | 8627.22 | 8589.24 | -37.99 | 1.00 |
| 5 | 1,031,390 | 9282.35 | 9343.00 | 60.65 | 0.99 |
| 6 | 1,031,435 | 10098.20 | 10236.47 | 138.27 | 0.99 |
| 7 | 1,031,413 | 11135.70 | 11333.84 | 198.14 | 0.98 |
| 8 | 1,031,461 | 12443.52 | 12777.93 | 334.41 | 0.97 |
| 9 | 1,031,389 | 14366.97 | 14938.86 | 571.89 | 0.96 |
| 10 | 1,031,425 | 20923.21 | 20125.01 | -798.20 | 1.04 |

The prediction model clearly misses the mark with the highest cost HHAs when compared with the second highest cost group of HHAs.

The following commentary seems to indicate that the prediction model may take into consideration "beneficiary characteristics that are outside the provider's control" (not previously reported/discussed in the submission):

The distribution of facility-level observed and risk-adjusted spending is shown in  and **.** By taking into account beneficiary characteristics that are outside the provider's control, the model compresses the distribution of provider-level spending and decreases their variability. The degree of compression demonstrates that there exists a significant amount of variation in HHA spending that is not explained by the observed beneficiary risk factors.

**Table 1. Distribution of Provider-Level Observed and Risk-Adjusted Episode Spending**

| Group | K | Mean | SD | 10th Pct | 25th Pct | 50th Pct | 75th Pct | 90th Pct |
|---|---|---|---|---|---|---|---|---|
| Observed | 10,470 | 10,295.8 | 1,959.4 | 7,570.4 | 9,069.1 | 10,477.8 | 11,586.1 | 12,525.3 |
| Predicted | 10,470 | 10,648.6 | 1,052.0 | 9,295.9 | 9,896.7 | 10,666.1 | 11,371.4 | 11,949.8 |

Analysis of Medicare Claims File for HH CY 2016-2017.

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

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

**Panel Member #1** Two approaches adopted: Statistical risk model and stratification

*Statistical risk model*

The MSPB-PAC HH risk adjustment models are adapted from the model used in the NQF-endorsed hospital MSPB measure (#2158), which is itself an adaptation of the standard CMS-HCC risk adjustment model. Risk adjustment was performed separately for the MSPB-PAC HH episode types listed below:

- HH Standard
- HH Low Utilization Payment Adjustment (LUPA)
- HH Partial Episode Payment (PEP)

The overall adjusted R-squared is 0.092. The overall adjusted R-squared was computed by fully interacting episode type (LUPA, PEP, or standard) with all other coefficients.

*Stratification*

Stratification entailed investigating Low Utilization Payment Adjustment (LUPA) and Partial Episode Payment (PEP) episodes, which represent adjustments to the standard HH payment policy and generally indicate non-standard circumstances. The measure developer tested both controlling for LUPA and PEP episode types and stratifying the sample and found that stratification resulted in improved overall model fit and sufficient sample sizes in all three strata (standard, LUPA, and PEP episodes). The adjusted R-squared for the HH Standard episodes is 0.090; it is 0.096 for HH LUPA episodes and 0.076 for HH PEP episodes.

The episode-level predicted costs range from $2,562 to $69,577, indicating that this model has a range of predictions and can predict both high and low costs. The model discrimination and calibration results demonstrate good predictive ability across the full range of episodes, from low to high spending risk (decile analyses).

Nevertheles, the R-squared of 0.1 (rounded) seems low for the risk adjustment models with such large samples and so many adjustors. It seems to suggest that models may not have incorporated some key predictors of costs, which may not have been observed in the data.

**Panel Member #2** In spite of making a strong conceptual argument for including SES, and statistical results showing spending was slightly higher for patients who were dual eligible ($11,926 vs $11,621 for non duals for the full population, or **2.6% higher** which can be a significant difference in the HH Medicare payment world, non-white (e.g., $12,060 for Blacks compared to $11,739 for whites, or **2.7% higher** which can have very significant impact on a plan with many non-white members, and had low SES based on AHRQ SES index ($11,962 in Q2 of SES Index vs $11,553 in highest quantile, or **3.5% higher**), and that each of those social risk factors was statistically significant in the risk adjustment model, they concluded that adding them, individually or together, did not substantially improve overall model fit. This is inconsistent with their argument that the SES factors are significant in the model due to large sample size. I would similarly argue that the small impact on R2 and model fit is NOT surprising given there were already **>120 variables** in the model, and patients with above SES characteristics tend to have more clinical risk factors as well; however the SES remained significant in the model. Further, they found including SES had minimal impact on average measure scores, so they chose to NOT adjust the scores for these social risk factors. Given the empirical findings of 2.6% to 3.5% higher costs for patients who are dual eligible, non-white and with low SES, which can be highly significant for HHAs operating on tight margins, I strongly believe the factors should have been kept in the model. In cases where an HHA serves a large proportion of patients with these SES factors, they will be penalized for having higher costs than expected as these factors are not accounted for, which is not intent of the measure and could restrict access to high quality HOME HEALTH for these patients.

The model as specified does have good discrimination properties based on clinical risk adjustments applied, though **the adjusted R-squared is very low 0.092.**

**Panel Member #4** Risk adjustment approach is standard approach for CMS measures. It appears to take into account underlying differences in disease/chronic conditions/age.

Social risk factors analyzed, and as implemented in risk adjustment model, make small-negligible difference in score or adjusted R-square. I have concerns about use of zip code level SES index, rather than census tract/address based adjustment, and we have no data on how much variability across facilities there is in social risk factors, and whether facilities with high proportion of disadvantaged SES patients regularly score worse, and whether the differences in scores are larger for these facilities.

The risk adjuster explains a small proportion of variance in cost. This is not unusual with cost measures. But the risk adjustment model is pretty much a cookie cutter of the standard risk adjustment model for CMS cost measures, and I don't see any thought given to what is unique about patients in HH, and what are the drivers of differences in costs that are not under the control of the HHA providers, to assure these are controlled for. There is also boiler plate language about not wanting to include too many interactions because of risk of overfitting in a multimillion record data set where the potentially interacted measures and cases with both or neither conditions are probably with adequate numbers that overfitting is not a problem. The key line is consistency with the standard CMS approach when we should be asking whether the risk adjustment model is appropriate for controlling for cost and use differences across HHAs that are not under the control of HHA.

**Panel Member #7** The overall R-squared was 0.092. The observed to predicted costs varied between 0.96 and 1.05 for the ten deciles.

**Panel Member #8** Concern about the lack of fit in the lowest decile (Table 11) **–** The same issue is not present in the highest decile; this raises the concern re: bias for that lower decile of spend.

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

**For cost/resource use measures ONLY:**

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

    XxX **Yes** X☐x **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

    Panel Member #3 Because the MSPB measure is a ratio, the magnitude of this value for HHAs may be over-estimated when compared with other post-acute care settings due to the relatively low cost of home-based care compared with the institutional care associated with IRFs, LTCHs, and SNFs.

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

    **Panel Member #1** None

    **Panel Member #3** If the MSPB HHA measure is used (see cavaet regarding data source cited), then the measure can only be used for comparison among HHAs and CANNOT be used to compare across other institutionally-based Provider types. Note: This may not be consistent with the stated IMPACT intent to create cross-Provider-type comparative metrics.

    **Panel Member #4** I think the committee should discuss the decision to allow overlapping episodes and treat the cost of care in second HHA as cost in both episodes.

    **Panel Member #8** General concern about the potential overlap in the various MSPB metrics submitted. It appears that each services may be attributed to multiple 'episodes' and providers. Unclear how improvements may be made when the attribution is so scattered.

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

### VALIDITY: TESTING

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

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

    ☐ **Face validity**

    ☒ **Empirical validity testing of the measure score**

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

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

    **Submission document: Testing attachment, section 2b2.2**

    **Panel Member #1** Empirical validity of the MSPB-PAC HH measure score was assessed through the following correlations that are conceptually expected to behave certain ways:

    (i) Comparison of the ratio of observed over expected spending for MSPB-PAC HH episodes with and without hospital admissions, and with and without emergency admission occurring in the episode period with the expectation that the variation in service utilization is captured by the MSPB-PAC cost measure.

    (ii) Correlation between MSPB-PAC HH scores and the Discharge to Community (DTC) rates and Acute Care Hospitalizations (ACH) for CY 2016-2017. The expectation is that MSPB-PAC HH and DTC will be negatively correlated because efficient providers (lower MSPB-PAC SNF scores) should have higher rates of successful discharge to the community while MSPB-PAC HH will be positively correlated indicating that less efficient hospitals have higher rates of unplanned hospital admissions.

(iii) Evidence of minimal correlation between MSPB-PAC HH scores and provider's functional improvement scores on ambulation, bathing, bed transfer, management of oral medications, and pain interfering with activity.

**Panel Member #2** The developers used the following methods to test reliability:

1. Evaluated correlation with known indicators of resource or service utilization (hospital admissions and emergency room (ER) visits during the episode period). They compared the ratio of observed over expected spending for MSPB-PAC HOME HEALTH episodes with and without hospital admissions occurring in the episode period. They also compared the observed over expected spending for episodes with and without ER visits. This analysis tested whether variation in service utilization is captured by the MSPB-PAC cost measure.

2. Examined the correlation between MSPB-PAC HOME HEALTH scores and the Discharge to Community (DTC) rates for HHAs (measure endorsed by NQF (#3477) is publicly reported and also based on Medicare claims.

3. Examined the correlation between MSPB-PAC HOME HEALTH scores and provider's scores on the Acute Care Hospitalization During the First 60 Days of Home Health are endorsed by NQF (#0171).

The developers hypothesized there would be a negative association between the MSPB measure and DTC measure, indicating that more efficient providers would have higher rates of successful discharge to the community. Conversely, they hypothesized that there would be a positive correlation between MSPB and ACH, indicating that less efficient providers have higher rates of unplanned hospital admissions. Providers whose patients have adverse events, such as re-hospitalization, at a rate higher than would be expected based on patient characteristics, should have lower DTC scores, higher ACH scores, and higher (less cost efficient) MSPB scores.

4. Examined the correlation between MSPB-PAC HH scores and provider's functional improvement quality measure scores publicly reported on the Home Health Compare website for CY 2017. Specifically, we use the following NQF-endorsed assessment-based measures:

- Improvement in ambulation (#0167)
- Improvement in bathing (#0174)
- Improvement in bed transfer (#0175)
- Improvement in management of oral medications (#0176)
- Improvement in pain interfering with activity (#0177)

These measures report risk-standardized proportions of episodes where patients' assessment scores indicate improvement between start of care and discharge. All are calculated using one year of data.

**Panel Member #3**

- The reference to the measure including the period "31 days after discharge" is consistent with the LTCH measure but not the IRF measure. Why?
- Comparison to utilization outcome (hospitalization and/or ED use) and functional outcomes is appropriate. Why not physiological outcomes such as pressure ulcers and UTI's?

**Panel Member #4** Medical expert review at contractor, CMS and TEP of exclusion of unrelated expenses may be valid, but process, detailed in Appendix D, not available for review. I cannot assess how systematic the review was, or how items were identified for consideration. The exclusions discussed in the documentation are reasonable, but I cannot be sure that all potentially excludable expenses were considered and evaluated.

For empirical validation, the submission included two tests:

--correlation with other related measures. Correlation appears reasonable.

--episodes with high cost components, such as readmissions to hospital, had higher than predicted expenses and high ER use. We have accepted this method of validation in prior measures, but I am concerned about the validity of including these costs in the model to begin with.

While the sponsors have not presented a case for face validity for this measure, and are not required to, I have concerns that the measure on its face is not valid.

**Panel Member #7** The MSPB-PAC HHA were examined for correlation with known indicators of resource or service utilization, specifically hospital admissions and emergency room visits. In addition, discharge to

rates were examined. Also, correlation was examined with quality measures for the HHA, specifically improvements in ambulation, bathing, bed transfer, management of oral medications, and pain interfering with activity. The latter measures had minimal correlation and DTC referral was negative.

**Panel Member #8** Correlation with other measures -

**Panel Member #9** The developer conducted three separate empirical tests on validity: 1) Assess how this measure is related to resource utilization, such as hospitalization within the episode window and,ER visit within the episode window. 2) Correlate this measure with the discharge to community rates measure and the acute care hospitalization rates measure. 3) Correlated this measure with a set of functional improvement quality measures. These are not unreasonable, but both hospitalization and ER visit are functionally related to the measure. And DTC. is also inherently related to this measure.

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

**Submission document: Testing attachment, section 2b2.3**

**Panel Member #1** The mean observed to expected cost ratio for episodes without a hospital admission is 0.68, compared with 2.31 for episodes with at least one hospital admission during the episode period (p-value<0.0001). The mean observed to expected cost ratio for episodes without an ER visit is 0.89, compared to 1.39 for episodes with at least one ER visits (p-value<0.0001).

Small but significant negative association between MSPB and DTC measure scores (-0.25), and positive MSPB and ACH (0.305).

Small, significant positive correlations (both Pearson and Spearman) between MSPB measure scores and functional improvement measure scores were found.

**Panel Member #2** The developers found a positive relationship between MSPB and known indicators of resource or service utilization. The mean observed to expected cost ratio for episodes without a hospital admission is 0.68, compared with 2.31 for episodes with at least one hospital admission during the episode period (p-value<0.0001). The mean observed to expected cost ratio for episodes without an ER visit is 0.89, compared to 1.39 for episodes with at least one ER visits (p-value<0.0001). They also observed a positive relationship between the mean observed to expected cost ratio and the number of hospitalizations/ER visits as hypothesized.

They also found a small but significant negative association between MSPB measure scores and the DTC measure scores as hypothesized and a very small but statistically significant correlation (Pearson -0.240; Spearman -0.250) between MSPB measure scores and DTC measure scores.

They found a small positive correlation (both Pearson and Spearman) between MSPB measure scores and ACH scores (Pearson 0.298; Spearman 0.305) which may indicate hospitalizations are associated with somewhat higher spending in 30 days after discharge from HOME HEALTH .

Finally, they found very small but significant positive correlations between MSPB scores and the various functional improvement scores as hypothesized (Pearson correlations ranging from 0.075 to 0.163;

Spearman ranged from 0.041 to 0.152), which may indicate therapy services (higher cost) are associated with somewhat higher rates of functional improvement on average.

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

**Panel Member #3**

- Relationship to utilization outcomes is in the direction expected—and values across comparison units (e.g., # of hospitalizations) are very large (0.68 for 0 hospitalizations vs. 3.41 for 4 or more hospitalizations) when compared with related MSPB measure values for IRF, LTCH, and SNF. This would indicate that comparison of MSPB values across Provider types is not valid. One reason for this comparatively large range of MSPB values for HHAs is the proportionally smaller actual costs (base value) of treating a patient in the home vs. in institutional setting (IRF, LTCH, and SNF). Hence, the impact of one or more hospital stays post-discharge is much larger for HHAs than for these other Provider types.

**Table 5. Mean Cost Ratio, by Number of Hospitalizations/ER Visits**

| Number of High-Cost Event | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Hospitalizations | 0.68 | 2.11 | 2.84 | 3.20 | 3.41 |
| ER Visits | 0.89 | 1.30 | 1.57 | 1.73 | 1.87 |

Analysis of Medicare Claims File for HH CY 2016-2017.

We also found a small, significant negative association between MSPB measure scores and the DTC measure scores (Table 6). Both Pearson and Spearman rank correlations revealed similar relationships.

**Table 6. Correlations between MSPB, Discharge to Community (DTC), and Acute Care Hospitalization (ACH) Measures**

| Measure Name | K* | Pearson Correlation | p-value | Spearman Correlation | p-value |
|---|---|---|---|---|---|
| Discharge to Community (DTC) | 9,711 | -0.240 | <0.001 | -0.250 | <0.001 |
| Acute Care Hospitalization (ACH) | 8,314 | 0.298 | <0.001 | 0.305 | <0.001 |

Analysis of Medicare Claims File for HH CY 2016-2017

If the MSPB HHA measure is used (see cavaet regarding data source cited), then the measure can only be used for comparison among HHAs and CANNOT be used to compare across other institutionally-based Provider types. Note: This may not be consistent with the stated IMPACT intent to create cross-Provider-type comparative metrics.

The interpretation of the relationship between functional outcome success and MSPB score is probably that "there is an increased cost due to the use of therapy (OT and PT) to produce these functional outcomes. Without the added costs, the likelihood of improvement in functional outcomes is decreased."

Lastly, we found very small, significant positive correlations (both Pearson and Spearman) between MSPB measure scores and functional improvement measure scores (Table 7). These results are consistent with the fact that functional improvement is not associated with lower rates of adverse

events; they may indicate that provision of additional therapy services is associated with somewhat higher rates of functional improvement, on average.

**Table 2. Correlations between MSPB and Functional Improvement Measures**

| Measure Name | K* | Pearson Correlation | p-value | Spearman Correlation | p-value |
|---|---|---|---|---|---|
| How often patients got better at walking or moving around | 8,646 | 0.128 | <0.0001 | 0.125 | <.0001 |
| How often patients got better at bathing | 8,662 | 0.163 | <.0001 | 0.150 | <.0001 |
| How often patients got better at getting in and out of bed | 8,573 | 0.153 | <.0001 | 0.152 | <.0001 |
| How often patients got better at taking their drugs correctly by mouth | 8,429 | 0.141 | <.0001 | 0.134 | <.0001 |
| How often patients had less pain when moving around | 8,572 | 0.075 | <.0001 | 0.041 | 0.0001 |

Analysis of Medicare Claims File for HH CY 2016-2017 and HH Compare data CY 2017.

**Panel Member #4** Empirical validation was consistent with past reviews but limited and incomplete.

I have two main problems with this measure:

First, while the concept is clear (all Part A and B Medicare expenses during the SNF visit and 30 day window after), there is no extended discussion of services actually included, and the major sources of variation in expenses across episodes. Some are specifically discussed, such as hospital readmission and ED, but I would like to better understand what differences in spending are systematically associated with higher or lower costs across patients and facilities to better assess whether these are under the control of the HHA or influenced by the care the HHA provides. I would also like to see more empirical testing of other sources of variation.

**I am concerned there are real issues of face validity of this measure.** The underlying issue for these measures is whether the care decisions made by the provider, in this case the HHA, contribute to the costs of care. There are two pathways for this: "profligate spending," i.e., high spending on a variety of services with low value for patients that run up cost with little gains in patient health or comfort, and complications to which the provider contributes that require treatment and add to cost. There is no discussion of the role of HHAs in preventing or contributing either to hospital admissions or ED use or any other high cost items that run up the expense during the period of treatment and 60 days after treatment ends for Standard Episodes and LUPA episodes. There is a certain element of cookie cutter measure development reflected in this measure -- 60 days after, standard risk adjuster—rather than tailoring to the service so that real cost variances associated with how the service is delivered are identified.

Second, at least in the materials provided, the extent and degree of systematic review of unrelated expenses to be excluded is not presented. Examples look reasonable, but full scope cannot be assessed. Nor is the method for identifying expenses to be considered for exclusion described. I can't assess whether there was a system or if it was ad hoc based on cases raised by various parties. If ad hoc, method is inadequate.

**Panel member #7** The measure was tested for relationship to known and imputed factors, and the results provided.

**Panel Member #8** Relationship with pressure ulcers weak – others are good.

**Panel Member #9** Given the nature of their relationship, the results for test 1 and 2 are as expected. The results from test 3 are very interesting, calling into question how this measure score should be used.

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

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

    ☐☒ **Yes**

    ☒ **No** ☐☐☒ **No Panel Member #3** (see comment about data source for HHA clinical information)☐

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

    See note above re basis for identifying and deciding to exclude unrelated expenses.

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 rationale in 21 and 22.

    **Panel Member #2** Validity results were based on several approaches and all showed strong validity based on hypothesized relationships.

    **Panel Member #3** Given the data source problem identified previously, I rated the validity as "insufficient." If the data source problem is addressed satisfactorily, then the validity rating would be changed to "low" given the wide variation in MSPB measure scores for HHAs and the potential use to compare MSPB costs for HHAs and other post-acute care settings (i.e., IRFs, LTCHs, and SNFs). Additionally, the distribution across the "below/at/above" the national MSPB value rating system presented is wildly different from the actual distribution of data presented in Figure 3.

    **Panel Member #4** Empirical validation was consistent with past reviews but limited and incomplete.

    Per response to q22:

Empirical validation was consistent with past reviews but limited and incomplete.

I have two main problems with this measure:

First, while the concept is clear (all Part A and B Medicare expenses during the SNF visit and 30 day window after), there is no extended discussion of services actually included, and the major sources of variation in expenses across episodes. Some are specifically discussed, such as hospital readmission and ED, but I would like to better understand what differences in spending are systematically associated with higher or lower costs across patients and facilities to better assess whether these are under the control of the HHA or influenced by the care the HHA provides. I would also like to see more empirical testing of other sources of variation.

I am concerned there are real issues of face validity of this measure. The underlying issue for these measures is whether the care decisions made by the provider, in this case the HHA, contribute to the costs of care. There are two pathways for this: "profligate spending," i.e., high spending on a variety of services with low value for patients that run up cost with little gains in patient health or comfort, and complications to which the provider contributes that require treatment and add to cost. There is no discussion of the role of HHAs in preventing or contributing either to hospital admissions or ED use or any other high cost items that run up the expense during the period of treatment and 60 days after treatment ends for Standard Episodes and LUPA episodes. There is a certain element of cookie cutter measure development reflected in this measure -- 60 days after, standard risk adjuster—rather than tailoring to the service so that real cost variances associated with how the service is delivered are identified.

Second, at least in the materials provided, the extent and degree of systematic review of unrelated expenses to be excluded is not presented. Examples look reasonable, but full scope cannot be assessed. Nor is the method for identifying expenses to be considered for exclusion described. I can't assess whether there was a system or if it was ad hoc based on cases raised by various parties. If ad hoc, method is inadequate.

**Panel Member #7** The measure was shown to correlate with other predictors of higher costs, namely hospital admission and EC visits.

**Panel Member #8** Correlation with related measures is strong.

**Panel Member #9** Outlier exclusion is a key concern. I would defer to the standing committee in terms of validity of this measure given the consistent positive correlation between MSPB and functional improvement measures.

### FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

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

### ADDITIONAL RECOMMENDATIONS

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

    **Panel Member #4** Concerns presented above.

**Panel Member #9** Outlier exclusion should be discussed further. Based on the measure specifications, episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution are excluded. Four factors make this approach particularly concerning: 1) This is a measure focusing on resource utilization, should episodes with very low or very high residuals be excluded? If the concern is with potentially undue influences of outliers, is exclusion the best available approach? 2) Winsorize predicted values: low predicted value below 0.5th percentile is already winsorized before calculation of residual. 3) Closing episodes: full payment for all claims that begin within the epidose window is counted toward the episode. This may make it likely that such episodes (with substantial claims at the end of episode window) become outleirs and be excluded. 4) There are substantial differneces between three different episode types, outlier exclusion based on residuals may favor one type of episode. At minimum, the developer should report the distribution of outlier exclusion across facilities to ensure that they don't concentrate in a limited number of facilities.

Additionally, I would defer to the standing committee in terms of validity of this measure given the consistent positive correlation between MSPB and functional improvement measures.

## 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 all data elements (e.g., DRG, ICD-9 codes on claims) are in defined fields in a combination of electronic sources, coded by someone other than person obtaining original information.
- The developer explained that the measure uses Medicare Enrollment data and Medicare FFS claims from the home health, inpatient, outpatient, and physician office settings claims data, which are routinely collected for payment purposes. Since these data are already collected as part of Medicare's payment process, this measure poses no additional data collection burden on providers.
- The developer also states that this measure uses data from the Minimum Data Set (MDS) which does not pose any additional burden on providers, as the submission of MDS is part of the federally mandated process for clinical assessment of all residents in Medicare and Medicaid certified nursing homes.

***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.**
- **Approach appears feasible**
- **No concerns with feasibility.**
- **all the data are available from secondary sources and the tools to construct the measure are in the public domain. This measure is constructed by CMS which has all the data.**
- **This measure is feasible to collect and calculate.**
- **Think the availability of social risk data that would improve risk adjustment not yet routinely captured**
- **No concerns**
- **Data elements routinely generated in the normal delivery of care and payment**

## 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 publicly reported as part of the Center of Medicare & Medicaid Services' Home Health Quality Reporting Program. (https://www.medicare.gov/homehealthcompare/search.html).
- The developer states that confidential feedback reports on the MSPB-PAC HH measure were provided to all active HH providers under the HH QRP starting in January 2018.
- Public reporting of the MSPB-PAC HH measure began in January 2019..

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

- The developer received comments from a range of stakeholders, including providers and provider associations, researchers, government agencies, information system vendors, advocacy groups, and the public during development and implementation.
- The comments received by the developer covered a range of topics, including episode construction, exclusions, score calculation, risk adjustment, and reporting. Several commenters commented on issues such as: usefulness of setting-specific MSPB-PAC measures, usefulness of a resource use measure as a measure of quality, the adequacy of the risk adjustment model, and the process of sharing measure scores with providers.
- The developer addressed all comments received by either revising the measure or by providing the rationale why revisions are not necessary or appropriate, before finalizing the measure in the CY 2017 HH PPS final rule.
    - Details of these considerations were provided in the public comment summary report.
    - The developer provided detailed descriptions of the systematic process used during development to identify clinically unrelated services, in response to public comments requesting more detail about the clinically unrelated excluded services.
    - In response to public comments about the inclusion of hospice services, the developer added a risk adjustor for when a hospice claim begins within the beneficiary's episode window.
    - The developer considered public comments about risk adjusting for prior hospital stays, aligning with other IMPACT Act measures, and added risk adjustors for length of prior inpatient and ICU stays.
    - Some commenters suggested controlling for community or family caregiver support. However, since HH is the only setting that has community or family caregiver support information available, inclusion of this adjustment would introduce inconsistencies between settings for the MSPB-PAC measures. Furthermore, testing indicated that the support information was 13% more likely to be unavailable.

**Additional Feedback:**

- The MAP PAC/LTC Workgroup met on December 14-15, 2015 and provided the preliminary decision of "encourage continued development" for this measure.
- The MAP Coordinating Committee considered these comments alongside the Workgroup recommendation and finalized the recommendation of "encourage continued development" in February 2016.

- The members found importance in balancing cost measures with quality and access, even though there were concerns about the ability to make comparisons across providers and premature discharges.
- Members noted the need to consider:
  - risk adjustment for severity and socioeconomic status. They urged CMS to incorporate functional status assessments into risk adjustment models to promote improvements.
  - the finalization of specifications to ensure costs are not double-counted between care settings; and recommended submission to NQF for endorsement. It was noted that the measures double count costs between providers and is inconsistent with IMPACT act to develop comparable resource measures of PAC providers.
- The MAP noted the measure was never fully specified before the PAC/LTC workgroup deliberations and the current specifications were released in mid-January with public comment period closing Jan 27th.
- Several members were uncomfortable with the MAP's final decision to recommend continued development.

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*

**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).

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

---

**4b1. Improvement results**

- This measure is being considered for initial endorsement.

**4b2. Unintended consequences**

- The developer states that there are no unexpected findings during the development and testing for the measure. However, they are aware of the need to continuously monitor for unintended impacts on patients, such as cost reduction at the expense of quality of care or avoiding complex, high

costpatients. Consequently, the developer plans to monitor trends in process and patient outcome measures, as well as trends in patient case-mix.

**4b2.Potential harms**

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

**4b3. Transparency**

- The developer states that the measure is publicly reported using routinely collected data from Medicare Enrollment data and Medicare FFS claims (from the home health, inpatient, outpatient, and physician office settings claims data) and the Minimum Data Set (MDS), which does not pose any additional burden on providers.
- Measure specifications and methods are transparent and available for users.

***Questions for the Committee**:*

- *What benefits, potential harms or unintended consequences should be considered?*

**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:**

- **Current use is public reporting for accountability, not yet used for payment.**
- **Unclear**
- **No additional comments beyond what NQF has already stated.**
- **The data were initially reported confidentially in 2017 and are publicly reported as of 2018. Implementation is already in play.**
- **This measure has been publicly reported since Fall 2018.**
- **yes**
- **No concerns**
- **Publicly reported**

**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**
- 
- **None.**
- **yes, there was a public comment period as part of the draft regulations. The feedback was considered and measure developer made some of the changes suggested.**
- **Although the measure has been publicly reported since Fall 2018, since this is the measure's initial endorsement, no improvement data was provided. It is unclear whether public reporting has provided HHAs with sufficient information to reduce costs.**
- **yes**
- **No concerns**
- **Feedback if provided to institutions**

**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:**

- **Confidential feedback reports include the following data, for the provider and for the national average: reporting period, number of eligible episodes, spending during treatment period, spending during associated services period, total spending during episode, average risk-adjusted spending, national median risk-adjusted spending, and the MSPB-PAC HH score. Not clear if this is sufficient to identify sources of high cost in an actionable way.**
- 
- **Yes, potential for improvement was demonstrated by incentivizing high-cost HHAs to the national mean.**
- **This is overall summary measure that provides directionality; more drill down information would help HHAs with changing practices that could reduce spend. So useability is medium grade.**
- **Although the measure has been publicly reported since Fall 2018, since this is the measure's initial endorsement, no improvement data was provided. It is unclear whether public reporting has impacted HHA costs.**
- **Seems to assume it can be used for improvement**
- **No concerns**

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

**Comments:**

- **No obvious harms.**
- **None.**
- **no harms at this stage, but always good to monitor over time.**
- **The measure could be more beneficial is the developers explored appropriate resource use and its relationship to health outcomes.**
- **none**
- **No concerns**
- **None identified**

## Criterion 5: Related and Competing Measures

- There are no competing measures for measure #3564 (i.e. same measure focus and target population)

- The developer identified the following NQF endorsed measures as related measures (same measure focus and approach but different target population as #3564):
    - 2158 : Medicare Spending Per Beneficiary (MSPB) – Hospital

**Harmonization**

- The developer states that MSPB-PAC measures are harmonized across PAC settings as well as with MSPB-Hospital. These measures were conceptualized uniformly across the four PAC settings in terms of the construction logic, the approach to risk adjustment, and measure calculation, to meet the mandate of the IMPACT Act. They express that this alignment of the MSPB-Hospital and MSPB-PAC measures creates continuous accountability and aligns incentives to improve care planning and coordination across inpatient and PAC settings.

**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:**

- **Part of a complementary set of PAC cost measures. Need to consider how SNF care is used in continuum of care associated with the hospitalization measure and other PAC measures.**

- **No competing measure identified. Related NQF-endorsed measure identified by the developer is: 2158: Medicare Spending Per Beneficiary (MSPB) – Hospital. The developer stated that the MSPB-PAC measures are harmonized across post-acute care (PAC) settings (Inpatient Rehabilitation Facility, Long Term Care Hospital, Skilled Nursing Facility and Home Health Agency) as well as with MSPB-Hospital.**

- **related to other MSPB measures. None competing.**

- **none.**

- **Agree with SMP commenter that there is potential overlap in the various MSPB metrics submitted. It appears that each service may be attributed to multiple 'episodes' and providers. Unclear how improvements may be made when the attribution is so scattered. Developer believes the PAC MSPB measures are in alignment with hospital MSBP measure.**

- **No concerns**

- **NA**

## Public and Member Comments

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

- **There have been no public comments or support/non-support choices as of this date.**

## 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)**

MSPB-PAC HH QM was developed to address the resource use domain of the Improving Medicare Post-Acute Care Transformation Act of 2014 (IMPACT Act). As part of the IMPACT Act, MSPB-PAC aims to achieve interoperability, data exchange, and standardized measurement among post-acute providers. The mandated use of MSPB-PAC measures is intended to allow for a greater ability to measure resource use and efficiency of care to improve outcomes, as well as encourage all PAC providers towards aligned incentives and care coordination.

Differences in post-acute care payments are a key driver of variation in Medicare spending overall.[1,2] In addition, there have been a number of studies demonstrating significant variability in HH care and outcomes, links between provider characteristics and readmissions, and significant opportunities for improvement.[3,4,5] The cost and quality link is important, with this resource use measure playing an important role in discerning value of HH care.

The MSPB-PAC HH measure was adopted by CMS for the HH Quality Reporting Program (QRP) and finalized in the CY 2017 HH Prospective Payment System (PPS) Final Rule.[6] Public reporting for the measure began in Fall 2018 through the HH Compare website.

[1] Institute of Medicine. (2013). Variation in Health Care Spending Assessing Geographic Variation. (July)

[2] Kahn, E. N., Ellimoottil, C., Dupree, J. M., Park, P., & Ryan, A. M. (2018). Variation in payments for spine surgery episodes of care: Implications for episode-based bundled payment. Journal of Neurosurgery: Spine, 29(2), 214–219.

[3] Murtaugh C.M., Deb P., Zhu C., Peng T.R., Barrón Y., Shah S., Moore S.M., Bowles K.H., Kalman J., Feldman P.H., Siu A.L. (2017). Reducing Readmissions among Heart Failure Patients Discharged to Home Health Care: Effectiveness of Early and Intensive Nursing Services and Early Physician Follow-Up. Health Services Research. 52(4), 1445-1472.

[4] Lohman M.C., Cotton, B.P., Zagaria, A.B., Bao, Y., Greenberg, R.L., Fortuna, K.L., Bruce, M.L. ( 2017). Hospitalization Risk and Potentially Inappropriate Medications among Medicare Home Health Nursing Patients, Journal of General Internal Medicine. 32(12), 1301-1308.

[5] Institute of Medicine. (2013). Variation in Health Care Spending Assessing Geographic Variation. (July)

[6] Medicare Program; Calendar Year 2017 Home Health Prospective Payment System Update; Home Value-Based Purchasing Model; and Home Health Quality Reporting Requirements for Federal Register, Vol. 81, No. 213. https://www.govinfo.gov/content/pkg/FR-2016-11-03/pdf/2016-26290.pdf

**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.**

MSPB-PAC HH measure scores are reported publicly for all US providers paid under Medicare's HH Prospective Payment System (PPS) with 20 or more eligible episodes in the reporting period. There were a total of 10,470 HHAs with 20 or more episodes in CY 2016-2017. Their scores represent 10,321,802 patient episodes, after all exclusions were applied. The scores show a good deal of variability – the descriptive statistics are provided below.

MSPB-PAC HH score descriptive statistics:

Mean: 0.96

Standard Deviation: 0.15

Min: 0.31

Max: 2.44

Interquartile range (75th percentile minus 25th percentile): 0.18

Score Percentiles

10: 0.78

20: 0.85

30: 0.90

40: 0.94

50: 0.97

60: 1.00

70: 1.03

80: 1.08

90: 1.13

**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.**

Not applicable

**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.**

Not applicable

**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.**

Not applicable

**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.**

MSPB-PAC HH is intended to allow for a greater ability to measure resource use and efficiency of care to improve outcomes, as well as encourage all PAC providers towards aligned incentives and care coordination. The measure assesses Medicare spending by HHAs and other healthcare providers during an MSPB-PAC HH episode. An MSPB-PAC HH episode includes all Medicare Part A and Part B services with a start date in the episode window, except for a limited set of services that are not clinically related to the episode. The episode window is opened by a trigger event (i.e., admission to an HHA) and ends 30 days after the discharge from that HHA. The measure is calculated as the ratio of the payment-standardized, risk-adjusted MSPB-PAC Amount for each HHA divided by the episode-weighted median MSPB-PAC Amount across all HHAs. The MSPB-PAC Amount is the ratio of the observed episode spending to the expected episode spending, multiplied by the national average episode spending for all HHAs.

## 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):*

Home Care

**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://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/2016_04_06_mspb_pac_measure_specifications_for_rulemaking.pdf

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

Per episode

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

Facility

**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.)*

This measure is based on Medicare FFS administrative claims and uses data from the Medicare enrollment database and Minimum Data Set (MDS). The enrollment database provides information such as date of birth, date of death, sex, reasons for Medicare eligibility, periods of Part A and Part B coverage, and periods in the Medicare FFS program. The MDS is used to construct a risk adjustment variable, indicating beneficiaries who have been institutionalized for at least 90 days in a given year. The data elements from the Medicare FFS claims are those basic to the operation of the Medicare payment systems and include data such as date of admission, date of discharge, diagnoses, procedures, and revenue center codes. The Medicare FFS claims data files are used to identify Medicare services from HHAs and other settings (e.g., the inpatient setting) within the episode window. No data beyond the claims submitted in the normal course of business are required from providers for the calculation of this measure.

This measure submission is based on CY 2016-2017 data, which were the most recent data available at the time of our analyses. We used the data sources listed below to develop the analytic file for measure specification and testing:

- Medicare Fee-For-Services claims and enrollment data: We accessed inpatient, outpatient, carrier, skilled nursing facility, home health, durable medical equipment, and hospice claims through the Centers for Medicare & Medicaid Services (CMS) Common Working File (CWF). The data dictionary for all Medicare FFS claims, demographic, and enrollment data are available at: https://www.resdac.org/cms-data?tid%5B%5D=4931&tid_1%5B%5D=1&=Find+Data+Files. General information about the CWF is available at: https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Downloads/clm104c27.pdf.
- Minimum Data Set (MDS): Acumen obtains the MDS through the Quality Improvement and Evaluation System (QIES). The data dictionary for the MDS data is available at: https://www.resdac.org/cms-data/files/mds-3.0/data-documentation.

We used two mappings to group diagnosis and procedure codes for use in identifying clinical events, implementing exclusions and applying risk adjustment:

- Agency for Healthcare Research and Quality (AHRQ) Clinical Classifications Software (CCS) groupings for Services and Procedures: Software is available for download at: https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp
- CMS-Hierarchical Condition Category (HCC) mappings of ICD-9 and ICD-10 codes: We used the Version 22 CMS-HCC mapping, which is included in the software available at: https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html.

We used five additional data sources for measure testing purposes only and not for measure specification:

- 2017 American Community Survey (ACS) 5-year estimate: We used the ACS to obtain the ZIP Code Tabulation Area (ZCTA) level measures needed to compute the Agency for Healthcare Research and Quality (AHRQ) Socioeconomic Status (SES) index score for use in social risk factor testing. This information is downloadable at the US Census website: http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t.
- Rural-Urban Continuum Codes 2013: We used this data source to construct rural-urban identifiers for social risk factor testing. These codes include county FIPS indicators, which are then merged onto our episode file. More information on this data source can be found at: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/.
- Provider of Services Current Files (POS File): We used this data source to describe the characteristics of HHAs included in measure specification and testing, such as census region, ownership type, and rurality, as reported in Table 1. The POS file contains data on characteristics of hospitals and other types of healthcare facilities, including the name and address of the facility and the type of Medicare services the facility provides, among other information. The data are collected through the CMS Regional Offices. General information about

the POS Files is available at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Provider-of-Services/index.html.

- Home Health Compare data: We used this data source to examine the relationship between MSPB and assessment-based quality measures. The Home Health Compare data include publicly reported HH quality measures. The data are available at https://data.medicare.gov/data/home-health-compare.

- Common Medicare Environment (CME) database: We extracted patient-level dual eligibility information from the CME database for social risk factor testing. CMS has designated the CME database as the single, enterprise-wide authoritative source for Medicare beneficiary enrollment and demographic data. The CME database integrates and standardizes different types of beneficiary data from CMS legacy systems. The CME database receives information from the EDB and also contains additional information not available in the EDB. A description of the CME is available at: https://www.ccwdata.org/documents/10280/19002256/medicare-enrollment-impact-of-conversion-from-edb-to-cme.pdf.

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

<SamplingMethodologySpecificDataSourceAttachment nodeType="0" />

**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:

Please supply the username and password:

Attachment:

### *Code Table:*

URL:

Please supply the username and password:

Attachment: S_6_Code_Table_04_29.xlsx

**Construction Logic**

**S.7.1. Brief Description of Construction Logic**

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

The MSPB-PAC HH measure assesses Medicare spending by HHAs and other healthcare providers during an MSPB-PAC HH episode. An MSPB-PAC HH episode includes all Medicare Part A and Part B services with a start date in the episode window, except for a limited set of services that are not clinically related to the episode. The episode window is opened by a trigger event (i.e., admission to the HHA) and ends 30 days after the discharge from that HHA. The measure is calculated as the ratio of the payment-standardized, risk-adjusted MSPB-PAC Amount for each HHA divided by the episode-weighted median MSPB-PAC Amount across all HHAs. The MSPB-PAC Amount is the ratio of the observed episode spending to the expected episode spending, multiplied by the national average episode spending for all HHAs.

An MSPB-PAC HH measure score of less than 1 indicates that a given HHA's resource use is less than that of the national median HHA during a performance period. An MSPB-PAC HH measure score of greater than 1 indicates that a given HHA's resource use is more than that of the national median HHA during a 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.)*

MSPB-PAC HH episodes assess all Medicare Part A and Part B claims for services delivered to a beneficiary during the episode window, subject to exclusions for particular services that are clinically unrelated to PAC treatment. Constructing an MSPB-PAC HH episode involves the following steps:

Defining the episode trigger, episode window, treatment period, and associated services period;

Excluding certain services from the episode that are clinically unrelated to PAC treatment and closing the episode.

Episode Trigger. Each episode is opened by an episode trigger. In the HH setting, each claim triggers its own episode. Adjacent home health claims are not collapsed into one episode given the existence of many long sequences of consecutive home health claims lasting over 180 days. Patient characteristics and treatment regimens can change significantly during this time. Allowing each home health claim to trigger a new episode promotes the accuracy of predicted episode payments by using the most recent patient information for each claim in the risk adjustment model.

Episode Window. The episode window consists of a treatment period and an associated services period.

Treatment Period. The treatment period of an MSPB-PAC HH episode begins on the day of the trigger and ends after 60 days for Standard episodes and Low Utilization Payment Adjustment (LUPA) episodes, and at discharge for Partial Episode Payment (PEP) episodes. A Standard episode is triggered by a home health claim to which neither a LUPA nor PEP adjustment applies. A LUPA episode is triggered by a home health claim to which a LUPA adjustment applies, that is, when there are four or fewer visits in a home health claim. A PEP episode is triggered by a home health claim to which a PEP adjustment applies. A PEP is a pro-rated adjustment for shortened episodes as a result of patient discharge and readmission to the same provider within the same 60-day home health claim, or patient transfer to another HH provider with no common ownership within the same 60-day claim. If a patient is discharged to a hospital, SNF, or IRF, and readmitted to the same HHA within the 60-day claim, a PEP adjustment does not apply. A home health claim to which both a PEP and LUPA adjustment apply triggers an HH PEP episode.

The treatment period includes Medicare Part A and Part B services delivered to a beneficiary that are provided directly or could reasonably have been managed by the attributed HH provider, and that are related to the beneficiary's care plan. Treatment services occurring on the first day of MSPB-PAC HH episodes are subject to exclusions related to prior institutional care, including ambulance transport and durable medical equipment, prosthetics, orthotics, and supplies (DMEPOS) orders preceding the patient's admission to the HHA. Treatment services are also subject to exclusions for particular services that are clinically unrelated to PAC treatment, as described in section S.9.1, below.

Associated Services Period. The associated services period starts at the trigger event for each of the MSPB-PAC HH episodes, and ends 30 days after the end of the treatment period. The associated services period is the time during which all non-treatment services are counted towards the episode (associated services). Such services are associated with HH care but are not provided directly, or could not reasonably have been managed by the attributed provider. For instance, an associated service includes an acute inpatient hospital admission for a complication arising during or after HH treatment. The Medicare spending for all Part A and Part B services during the associated services period are counted toward the episode, with exceptions for clinically unrelated services, as described in section S.9.1, below.

Closing Episodes. MSPB-PAC HH episodes end 30 days after the end of the treatment period. The full payment for all claims that begin within the episode window is counted toward the episode; this is done to maintain consistency with the MSPB-Hospital measure (NQF #2158) and to fairly assign payment to the episode for Medicare claims paid on a prospective payment system, regardless of episode length.

An MSPB-PAC HH episode may begin during the associated services period of another MSPB-PAC HH episode in the 30 days post-treatment. See section S.7.3 for examples of situations in which this occurs and how it is handled in the MSPB-PAC HH measure.

Measure Calculation

Certain episodes are excluded from the MSPB-PAC HH calculation to ensure that the measure facilitates meaningful comparisons between HH providers. These exclusions are distinct from the exclusions for clinically unrelated services discussed above, which exclude a limited set of services from MSPB-PAC HH episodes. In contrast, episode-level exclusions, discussed in section S.9.1, remove entire episodes from measure calculation when certain criteria are met.

After applying the episode-level exclusions, the measure can be calculated in the following steps:

Step 1: Standardize Claim Payments. The first step in calculating the standardized payment for a claim is to eliminate variation in payments due to Medicare geographic adjustment factors and add-on payments for Medicare programs, such as indirect medical education (IME) and disproportionate share hospitals (DSH). The goal of this step is to remove sources of variation not directly related to decisions to provide clinical services. Payment standardization controls for geographic variation in healthcare payments, such as the hospital wage index and geographic practice cost index (GPCI).[1] All payment data shown in the MSPB-PAC HH measure and supporting documentation reflect allowed amounts, which include both Medicare trust fund payments and beneficiary deductible and coinsurance. Bonus or penalty amounts due to Medicare quality reporting or other special programs are not included.

Step 2: Calculate Standardized Episode Payments. Next, to prepare claims data for calculating risk-adjusted payments, standardized episode payments are calculated. For each episode, standardized payments include all standardized Medicare claims payments for services in the episode window, as detailed in previous paragraphs.

Step 3: Calculate Predicted Episode Payments. The third step calculates predicted payments for each episode. This step estimates the relationship between the independent variables and standardized episode payments using an ordinary least squares (OLS) regression. The calculation is performed separately for Standard, LUPA, and PEP episodes (described above). See Appendix C of the Measure Specifications document provided in section S.1 for a full list of the independent variables used in the risk adjustment model.[2]

Step 4: Winsorize (Bottom Code) Predicted Values. Next, the distribution of predicted values is examined. If the distribution of predicted values includes extremely low values (defined as below the 0.5th percentile), winsorization is performed at the low end of the distribution (i.e., “bottom coding”). The resultant values are renormalized to maintain a consistent average episode payment. If the distribution of predicted values does not include extremely low values, winsorization is not required to ensure meaningful ratios of observed to predicted spending (see below). In accordance with the MSPB-Hospital measure (NQF #2158) calculation, renormalization multiplies the winsorized predicted values by the ratio of the average original predicted payment and the average winsorized predicted payment. For example, suppose an episode’s predicted value (PREDICTED_VALUE) is $1,000, but the 0.5th percentile of predicted values is $1,500. Then, that episode’s “winsorized” predicted value (WINS_PREDICTED_VALUE) would be $1,500. The “renormalized” winsorized predicted value would be:

($1500×mean(PREDICTED_VALUE))/mean(WINS_PREDICTED_VALUE)

where the mean is taken over the entire national sample of the MSPB-PAC HH episodes. This re-normalization ensures that the average of the resulting winsorized predicted values is equal to the average of the original predicted values.

Step 5:

Calculate Residuals. The residuals for each episode are calculated as the difference between the standardized episode spending and the standardized predicted spending for episode i and HHA k.

Residual_ik = Y_ik - (Y_ik)^

where:

Y_ik is the attributed standardized spending for episode i and provider k.

(Y_ik)^ is the standardized predicted spending for episode i and provider k, as predicted from risk adjustment.

Step 6: Exclude Episodes with Outlier Residuals. The next step excludes outliers from the calculation and renormalizes the resultant predicted values to maintain a consistent average episode payment level. Episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution are excluded, reducing the impact of high- and low-payment outliers on a PAC provider's measure. Predicted values after outlier exclusion are renormalized by multiplying each value by the ratio of the average standardized un-risk-adjusted payments to the average of the standardized predicted payments remaining after exclusion of episodes with outlier residuals.

Step 7: Calculate MSPB-PAC HH Measure. The MSPB-PAC HH measure is calculated for individual providers, allowing them to be compared relative to other HH providers nationally. The MSPB-PAC HH measure is calculated as the ratio of the MSPB-PAC Amount for each HH provider divided by the episode-weighted median MSPB-PAC Amount across all HH providers. MSPB-PAC HH measure calculation is performed separately for HH Standard, LUPA, and PEP episodes to ensure that they are compared only to other episodes of the same type. The final MSPB-PAC HH measure combines the ratios of the episode types to construct one provider score.

To calculate the MSPB-PAC Amount for each HH provider, one calculates the ratio of the standardized spending for HH Standard episodes over the expected spending (as predicted in risk adjustment) for HH Standard episodes, the ratio of the standardized spending for HH LUPA episodes over the expected spending (as predicted in risk adjustment) for LUPA episodes, and the ratio of the standardized spending for HH PEP episodes over the expected spending (as predicted in risk adjustment) for PEP episodes, and then averages these ratios across all episodes for the attributed provider. This quantity is then multiplied by the average episode spending level across all HH providers nationally for Standard, LUPA, and PEP episodes.

Mathematically, MSPB-PAC HH for individual provider k is:

?"MSPB-PAC HH Amount" ?_k/"National Median MSPB-PAC HH Amount"

The numerator is the MSPB-PAC HH Amount, or the average risk-adjusted episode spending across all episodes for the attributed provider. This is then multiplied by the national average episode spending level for all HH providers nationally. Mathematically, the MSPB-PAC HH Amount numerator is calculated as:

"MSPB-PAC HH Amount"_k = ((1/n_k)*Sum_i_{i_k}[Y_ik/(Y_ik)^])*((1/N)*Sum_k[Sum_i_{i_k}[Y_ik]])

Where:

Y_ik is the attributed standardized spending for episode i and provider k.

(Y_ik)^ is the expected standardized spending for episode i and provider k, as predicted from risk adjustment, and resulting from Step 6 above

n_k is the number of episodes for provider k

N is the number of episodes nationally

i_{i_k } is all episodes i in the set of episodes attributed to provider k

The denominator is the episode-weighted national median of the MSPB-PAC HH Amounts for all HHAs nationally.

The MSPB-PAC HH measure is calculated for each provider. An MSPB-PAC HH measure score with a value less than 1 indicates that a given HHA's resource use is less, after risk-adjustment, than the resource use of the national median MSPB-PAC HH Amount across all HHAs nationally in the given performance period.

Notes:

[1] QualityNet, "CMS Price (Payment) Standardization Overview" (April 2019)) https://www.qualitynet.org/inpatient/measures/payment-standardization

[2] Centers for Medicare and Medicaid Services. Measure Specifications: Medicare Spending Per Beneficiary-Post-Acute Care Resource Use Measures. April 2016. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/2016_04_06_mspb_pac_measure_specifications_for_rulemaking.pdf

**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:

Please supply the username and password:

Attachment: Home_Health_MSPB_Section_S.7.2.docx

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

The MSPB-PAC HH measure methodology does not separate concurrent events. The MSPB-PAC HH measure methodology defines an MSPB episode as all claims from the first day of the HH claim to 30 days after the end of the treatment period. Please refer to section S.8.4., which details the rationale for the construction of the MSPB-PAC HH episode, for a discussion of the advantages of this approach.

The definition of MSPB-PAC HH episodes allow one episode to overlap with other episodes. One possible scenario occurs where an HH provider discharges a beneficiary who is then admitted to another HHA within 30 days. In this case, the second episode begins in the associated services period of the first episode in the 30 day post-treatment. The HH claim will be included once as an associated service for the attributed provider of the first MSPB-PAC HH episode and once as a treatment service for the attributed provider of the second MSPB-PAC HH episode. This overlap is necessary to ensure continuous accountability between providers throughout a beneficiary's trajectory of care, as both providers share incentives to deliver high quality care at a lower cost to Medicare and engage in patient-focused care planning and coordination.

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

We do not provide specifications for linking complementary services.

An MSPB-PAC HH episode includes all Medicare Part A and Part B services that fall within the episode window that starts at the first day of the HH claim and ends 30 days after the end of the treatment period, except for a limited set of services that are excluded for being clinically unrelated to HH treatment (as described in section S.9.1).

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

Clinical Classification Software (CCS). We use CCS for Services and Procedures to group HCPCS codes on outpatient (OP) claims that occur during the episode into clinically meaningful categories. This grouping is used

to identify clinically unrelated events which are then excluded from episode calculation (as detailed in section S.9.1).

Hierarchical Condition Categories (HCC). Hierarchical Condition Categories with a 90-day lookback period are included as covariates in the risk-adjustment model. The MSPB-PAC HH risk adjustment methodology is discussed in additional detail in section S.12.

Clinical Case-Mix Category. The clinical case-mix category variables used in the MSPB-PAC HH risk-adjustment model are included to account for beneficiary characteristics prior to the start of an MSPB-PAC HH episode that may influence the type and intensity of care. Taking the most recent institutional claim (by end date) in the 60 days prior to the start of an MSPB-PAC HH episode, the episode is assigned to one of the following mutually exclusive and exhaustive clinical case-mix categories:

(1) Prior Acute Surgical IP – Orthopedic – beneficiaries who have most recently undergone orthopedic surgery in an acute inpatient hospital

(2) Prior Acute Surgical IP – Non-Orthopedic – beneficiaries who have most recently undergone a non-orthopedic surgery in an acute inpatient hospital

(3) Prior Acute Medical IP with Intensive Care Unit (ICU) – beneficiaries who have most recently stayed in an acute inpatient hospital for non-surgical reasons and had a stay in the ICU

(4) Prior Acute Medical IP without ICU – beneficiaries who have most recently stayed in an acute inpatient hospital for non-surgical reasons but did not have a stay in the ICU

(5) Prior PAC - Institutional – beneficiaries who are continuing PAC from an institutional PAC setting (i.e., coming from an LTCH, IRF, or SNF)

(6) Prior PAC - HHA – beneficiaries who are continuing PAC from an HHA

(7) Community – all other beneficiaries

In the event that there are multiple prior claims with the same end date in the 60 days prior to the start of a PAC episode, additional logic is employed to determine the episodes' clinical case-mix category. For conflicts occurring between two IP claims, the clinical case-mix category corresponding to the claim with the longest length of stay (LOS) is assigned. For all other types of conflicts including those where the LOS is the same between two IP claims, the clinical case-mix category is assigned using a hierarchy in the order of the categories listed above.

**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)*

We do not provide measure specifications or guidelines for missing data :

Accurate and complete Part A and Part B claims are necessary for physician and hospital billing. Thus, missing data on Medicare enrollment and claims are very rare. All the data used to calculate MSPB-PAC HH measure values are included on Medicare claims and enrollment data. The data fields used to calculate the MSPB-PAC HH measure (e.g., payment amounts, DRGs, diagnosis and procedure codes, etc.) are included in all Medicare claims because HHAs only receive payments for complete claims. We have complete data for each beneficiary who has an MSPB-PAC HH episode, since beneficiaries are excluded if they are not continuously enrolled in only Medicare Parts A and B or if Medicare is not the primary payer during an episode, as described in section S.9.1. This ensures that we have all claims data for beneficiaries included in the MSPB-PAC HH measure calculation.

**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

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

Durable Medical Equipment (DME)

Other services not listed

All services covered by Medicare Part A and B (Hospice, SNF, Home Health, and services captured in carrier claims.)

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

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

The MSPB-PAC HH measure assesses the standardized allowed amounts of services performed by HHAs and other healthcare providers during an MSPB-PAC HH episode, which includes all Part A and Part B Medicare claims from the first day of the HH claim to 30 days after the end of the treatment period. As a result, costs from all Part A and Part B claim types (i.e., inpatient, outpatient, home health agency, hospice, skilled nursing facility, durable medical equipment, and carrier) are included. Note that costs of Part B drugs are included, but costs of Part D drugs are not included since Part D is not used to calculate the MSPB-PAC HH measure. The methodology used to standardize payment for these claims is available for download from the URL provided in section S.7.8a ("CMS Price (Payment) Standardization").

**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: https://qualitynet.org/inpatient/measures/payment-standardization

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 MSPB-PAC HH measure aims to calculate resource use in the period between the start of the treatment period and the end of the associated services period. The clinical topic area includes all HH claims in the United States. The measure accounts for differences in payment policy and beneficiaries' underlying health characteristics by stratifying by standard and site neutral payment rate admissions. To adjust for beneficiary characteristics that are out of the influence of the attributed HHA and may affect resource use, we risk-adjust the total observed episode spending (described in section S.12) using CMS-HCC indicators and interactions

between selected comorbidities. In addition to comorbidities, we also include indicators for clinical case-mix based on diagnosis and procedural information on the most recent institutional claim (by end date) in the 60 days prior to the start of an MSPB-PAC HH episode. The MSPB-PAC HH episode encompasses all procedures and clinical events that occur between the start of the treatment period and 30 days after the end of the treatment period.

**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.)*

In order to create a resource use measure that is clinically valid, multiple steps were involved in excluding the least clinically relevant codes. Using an episode window (first day of the HH claim to 30 days after the end of the treatment period), we organized claims into clinically meaningful service categories or settings. For example, Medicare Severity-Diagnosis Related Groups (MS-DRGs) noted after an HH discharge were evaluated as medical or surgical admissions post-discharge. Clinical Classifications Software (CCS) and Current Procedural Terminology/Healthcare Common Procedure Coding System (CPT/HCPCS) services were organized into outpatient services, emergency department (ED) services, and durable medical equipment claims and evaluated for their relevance or relatedness to HH care.

Extensive clinical review was performed by clinicians with experience and expertise providing care in HH settings, as well as in collaboration with Medical Officers at CMS. The inpatient, outpatient, Part B physician and supplier, and DMEPOS services least clinically related to the HH care were excluded from the measure. For instance, services related to the routine management of preexisting chronic conditions (e.g., dialysis for ESRD, treatment for preexisting cancers, and treatment for organ transplants) were deemed clinically unrelated to the scope of the type of care that HHAs provide. Therefore, these types of services were excluded. Services were excluded if there was consensus across clinicians from the measure developer, external clinical experts including TEP members, and CMS medical officers. Please see section S.9.1 for overall clinical consensus regarding the types of exclusions.

To account for the association between clinical severity and resource use, we risk adjust the total observed episode spending (described in section S.12) using CMS-HCC indicators and interactions between selected comorbidities. Diagnosis codes on claims that occur during the 90-day period prior to the start of an MSPB-PAC HH episode (90-day “look back”) are used to create HCC indicators. The MSPB-PAC HH measure accounts for comorbid conditions and interactions by broadly following the CMS-HCC risk adjustment methodology, which is derived from Medicare Part A and B claims and is used in the Medicare Advantage (MA) program. For example, the measure accounts for interactions between disability status and selected HCC groups (e.g., Cystic Fibrosis, Severe Hematological Disorders, Opportunistic Infections, among others). Given the fact that beneficiaries often have more than one comorbidity, the model also includes commonly observed paired condition interactions, (e.g., chronic obstructive pulmonary disease [COPD] and congestive heart failure [CHF]) and commonly observed triple-interactions (e.g., diabetes mellitus, congestive heart failure, and renal failure). The full list of variables used in the risk adjustment model can be found in the Measure Specifications document provided in section S.1.

In addition to comorbidities, the MSPB-PAC HH measure utilizes clinical case-mix categories to create clinically meaningful subgroups that influence the type of services a beneficiary will receive from an HHA. To create these subgroups, information was derived from the institutional claim of the most recent hospitalization. The clinical case-mix category variables used in the MSPB-PAC HH risk-adjustment model are included to account for differences in intensity and type of care received by beneficiaries prior to the start of an MSPB-PAC HH episode. Taking the most recent institutional claim (by end date) in the 60 days prior to the start of an MSPB-PAC HH episode, the episode is assigned to one of the following mutually exclusive and exhaustive clinical case-mix categories:

1) Prior Acute Surgical IP – Orthopedic – beneficiaries who have most recently undergone orthopedic surgery in an acute inpatient hospital

2) Prior Acute Surgical IP – Non-Orthopedic – beneficiaries who have most recently undergone a non-orthopedic surgery in an acute inpatient hospital

3) Prior Acute Medical IP with ICU – beneficiaries who have most recently stayed in an acute inpatient hospital for non-surgical reasons and had a stay in the ICU

4) Prior Acute Medical IP without ICU – beneficiaries who have most recently stayed in an acute inpatient hospital for non-surgical reasons but did not have a stay in the ICU

5) Prior PAC - Institutional – beneficiaries who are continuing PAC from an institutional PAC setting (i.e., coming from an LTCH, IRF, or SNF)

6) Prior PAC - HHA – beneficiaries who are continuing PAC from a HHA

7) Community – all other beneficiaries

To simplify the clinical logic and avoid the issue of attributing claims to MSPB-PAC HH episodes in the case of concurrent clinical events, all claims that begin within the episode window (treatment period and associated services period) are included in the MSPB-PAC HH measure. A new episode may begin during the associated services period of a previous MSPB-PAC HH episode in the 30 days after the end of the treatment period of that HHA.

**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)*

As part of the IMPACT Act, the goals of MSPB-PAC were to standardize assessment data to allow for interoperability, data exchange, and standardized measurement among post-acute providers. It also mandated the use of quality measures for PAC.[1] This would ultimately allow for greater ability to measure resource use and efficiency of care to improve outcomes, as well as encourage all PAC providers towards aligned incentives and care coordination.

There have been a number of studies demonstrating significant variability in home health care and outcomes, links between agency characteristics and readmissions, and significant opportunities for improvement.[2,3,4] The cost and quality link is important, with this resource use measure playing an important role in discerning value of HH care. Indeed, post-acute care is a significant cause of variation in Medicare spending.[5,6]

Within PAC, accounting for the beneficiary's clinical severity is crucial for accurately predicting resource use within the PAC episode. There is ample evidence in both inpatient and post-acute settings that Medicare episode payments are associated with clinical case-mix or severity. For example, Vertrees et al. (2013) tested the relationship between case-mix and Medicare payments for acute hospitalization episodes using MS-DRGs and Clinical Risk Groups (CRGS), which are similar to the HCC groupings.[7] This analysis found that Medicare costs of acute-hospitalization episodes that use 30, 60, and 90-days post-discharge windows can be predicted in part by the use of clinical severity groupings and case mix indicators as risk adjustors. Clinical severity groupings, such as those defined in section S.8.2, have been shown to be associated with resource use in the post-acute settings.

Within this framework, the MSPB-PAC HH measure was created to assess care provided by home health agencies. Prior literature has shown that patient severity is also an important consideration in this setting. Madigan et. al (2012) noted that there is significant variation of readmissions across HHAs. They found that HHAs that had high risk of readmission were noted to care for patients with a higher home health care visit intensity, an indicator of clinical acuity.[8] Accounting for clinical severity and case-mix is crucial to estimating the expected Medicare cost of the episode and thus supports the intent of the MSPB-PAC HH quality measure.

The MSPB-PAC HH measure methodology defines an MSPB-PAC episode as all claims with start dates falling between the treatment period start date (date of admission to the HHA) and the associated services period end date (30 days post HH discharge). This episode definition is consistent with NQF's theoretical definition of an episode of care in that it is "…a series of temporally contiguous healthcare services related to the treatment of a given spell of illness or provided in response to a specific request by the patient or other relevant entity."[9] Moreover, NQF has endorsed multiple episode-based measure of resource use and cost, including the MSPB-Hospital Measure (NQF #2158) and hospital-level measures for episodes of care for pneumonia (NQF #2579), acute myocardial infarction (NQF #2431), and heart failure (NWF #2436). Each of these measures estimate the risk-adjusted cost of a hospital based episode of care covering 30 days post-admission or post-discharge.

Notes:

[1] Medicare Program; Inpatient Rehabilitation Facility Prospective Payment System for Federal Fiscal Year 2017 Federal Register, Vol. 81, No. 151. https://www.gpo.gov/fdsys/pkg/FR-2016-08-05/pdf/2016-18196.pdf

[2] Chen H.F., Carlson E., Popoola T., Suzuki S. (2016). The Impact of Rurality on 30-Day Preventable Readmission, Illness Severity, and Risk of Mortality for Heart Failure Medicare Home Health Beneficiaries. Journal of Rural Health. 32(2), 176-87.

[3] Murtaugh C.M., Deb P., Zhu C., Peng T.R., Barrón Y., Shah S., Moore S.M., Bowles K.H., Kalman J., Feldman P.H., Siu A.L. (2017). Reducing Readmissions among Heart Failure Patients Discharged to Home Health Care: Effectiveness of Early and Intensive Nursing Services and Early Physician Follow-Up. Health Services Research. 52(4), 1445-1472.

[4] Lohman M.C., Cotton, B.P., Zagaria, A.B., Bao, Y., Greenberg, R.L., Fortuna, K.L., Bruce, M.L. ( 2017). Hospitalization Risk and Potentially Inappropriate Medications among Medicare Home Health Nursing Patients, Journal of General Internal Medicine. 32(12), 1301-1308.

[5] Institute of Medicine. (2013). Variation in Health Care Spending Assessing Geographic Variation. (July)

[6] Kahn, E. N., Ellimoottil, C., Dupree, J. M., Park, P., & Ryan, A. M. (2018). Variation in payments for spine surgery episodes of care: Implications for episode-based bundled payment. Journal of Neurosurgery: Spine, 29(2), 214–219.

[7] Vertrees, J. C., Richard, A. F., Eisenhandler, J, Quain, A., & Switalski, J. (2013). Bundling post-acute care services into MS-DRG payments. Medicare & Medicaid Research Review 3, no. 3

[8] Madigan E.A., Gordon N.H., Fortinsky R.H., Koroukian S.M., Piña I., Riggs J.S. (2012). Rehospitalization in a national population of home health care patients with heart failure. Health Services Research. 47(6), 2316-38.

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

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

URL:

Please supply the username and password:

Attachment: MSPB-PAC_HH_NQF_Testing_Form_-_Appendix_Tables.xlsx

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

MSPB-PAC HH episodes include services that take place during the 30 days after the end of the treatment period in order to emphasize the importance of care transitions and care coordination in improving post-acute care and reducing unnecessary service use. As a result, services between the first day of the HH claim and 30 days after the end of the treatment period are attributed to the HH episode. This timeframe was selected to align with other measures and was felt to be long enough to capture clinically relevant events, but not so long as to reduce the attributed facility's influence in future events.

The advantages of this measure trigger and end mechanism are twofold. First, this approach is simple and easily implementable since it includes all claims, except those described in section S.9.1, during the MSPB-PAC HH episode. Second, the MSPB-PAC HH approach incorporates costs due to care complications unrelated to the original reason for home health services, encouraging HH care coordination.

**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 model, as described in section S.12.

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

Co-morbidities and disease interactions are accounted for in the MSPB-PAC HH measure risk adjustment methodology, as discussed in sections S.8.2 and S.12. Conditions which most directly impact beneficiaries' health status at the start of HH services are captured in the risk-adjustment via the 90-day look-back period prior to the start of an episode. Because the relationship between comorbidities and episode cost may be non-linear in some cases (i.e., beneficiaries may have more than one disease during an HH episode), the model also takes into account a limited set of interactions between HCCs and/or enrollment status variables. Example variable interaction terms include Diabetes Mellitus/Congestive Heart Failure, Renal Failure/Congestive Heart Failure, and Disability/Opportunistic Infections (for a complete list of these variable interaction terms and other risk adjustment variables, please refer to Appendix C of the Measure Specifications document provided in section S.1). The MSPB-PAC HH measure risk adjustment methodology includes only a limited set of interaction terms for two reasons. First, inclusion of too many interaction terms will over-fit the model. Second, the MSPB-PAC HH measure risk adjustment methodology broadly follows the established CMS-HCC risk adjustment methodology, which uses similar interaction terms.

**Adjustments for Comparability**

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

:

Exclusion of clinically unrelated services. Certain services are excluded from the MSPB-PAC HH episodes because they are clinically unrelated to HH care and/or because HH providers may have limited influence over certain Medicare services delivered by other providers during the episode window. These limited service-level exclusions are not counted towards a given HHA's Medicare spending to ensure that beneficiaries with certain conditions and complex care needs receive the necessary care. The list of excluded services was developed by obtaining consensus on the exclusion of each service from CMS clinicians, eight independently contracted clinicians (including two TEP members) with expertise in each of the PAC settings, and the measure developer's clinicians. Feedback from the TEP provided through the in-person meeting and follow-up email survey was also taken into consideration. Additional information on the process for developing the list of clinically unrelated services is available in Appendix D of the Measure Specifications document provided in

section S.1. The specialties of the non-CMS clinicians with whom we consulted during the measure development process are provided in Appendix F of the Measure Specifications document provided in section S.1. Services that were determined by clinical consensus to be outside of the control of PAC providers include:

- Planned hospital admissions[1]
- Routine management of certain preexisting chronic conditions (e.g., dialysis for end-stage renal disease (ESRD), enzyme treatments for genetic conditions, treatment for preexisting cancers, and treatment for organ transplants)
- Some routine screening and health care maintenance (e.g., colonoscopy and mammograms)
- Immune modulating medications (e.g., immunosuppressants for organ transplant or rheumatoid arthritis)

Other Exclusions. Once clinically unrelated services are excluded at the claim line level, we exclude episodes based on several other characteristics, such as:

1) Any episode that results from a Request for Anticipated Payment (RAP)

Rationale: HHA requests for anticipated payment claims are interim claims that do not reflect the final payment made by Medicare for the services.

2) Any episode that is triggered by an HH claim outside the 50 states, D.C., Puerto Rico, and U.S. Territories.

Rationale: This exclusion ensures that complete claims data are available for each provider.

3) Any episode where the claim(s) constituting the attributed HH provider's treatment have a standard allowed amount of zero or where the standard allowed amount cannot be calculated.

Rationale: Episodes where the claim(s) constituting the attributed PAC provider's treatment are zero or have unknown allowed payment do not reflect the cost to Medicare. Including these episodes in the calculation of MSPB-PAC HH measure could potentially misrepresent a providers' resource use.

4) Any episode in which a patient is not enrolled in Medicare FFS for the entirety of a 90-day lookback period (i.e., a 90-day period prior to the episode trigger) plus episode window (including where a beneficiary dies) or is enrolled in Part C for any part of the lookback period plus episode window.

Rationale: Episodes meeting this criteria do not have complete claims information that is needed for risk-adjustment and the measure calculation as there may be other claims (e.g., for services provided under Medicare Advantage [Part C]) that we do not observe in the Medicare Part A and B claims data. Similarly, episodes in which the patient dies are, by definition, truncated episodes and do not have a complete episode window. Including these episodes in the MSPB-PAC HH measure could potentially misrepresent a provider's resource use. This exclusion also allows us to faithfully construct Hierarchical Condition Categories (HCCs) for each episode by scanning the lookback period prior to its start without missing claims.

5) Any episode in which a patient has a primary payer other than Medicare for any part of the 90-day lookback period plus episode window.

Rationale: When a patient has a primary payer other than Medicare, complete claims data may not be observable. These episodes are removed to ensure that the measures are accurately calculated using complete data.

6) Any episode where the claim(s) constituting the attributed HH provider's treatment include at least one related condition code indicating that it is not a prospective payment system bill.

Rationale: Claims that are not a prospective payment system bill may not report sufficient information to allow for payment standardization.

7) Any episode with problematic claims data (e.g., anomalous records for stays that overlap wholly or in part, or are otherwise erroneous or contradictory)

Rationale: The episode with the most recent processing date is kept to ensure the accuracy of data elements.

Finally, as part of the measure construction process described in section S.7.2, episodes with residuals below the 1st or above the 99th percentile of the residual distribution are excluded, reducing the impact of high- and low-payment outliers.

Notes:

[1] The lists of clinically unrelated services built off the planned readmissions algorithm developed by the Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation, as well as the expansions to the Yale algorithm by RTI. Clinicians reviewed the list of exclusions from that algorithm in the context of PAC treatment. During the review process, clinicians reviewed admissions observed in MSPB-PAC episodes and created exclusions that overlap with the Yale algorithm. Details on the Yale and RTI algorithms are available here: "Hospital-Wide All-Cause Unplanned Readmission Measure - Version 4.0," in 2015 Measure Updates and Specifications Report, ed. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (2015). 10-11. Laura Smith, West, S., Coots, L., Ingber, M., "Skilled Nursing Facility Readmission Measure (SNFRM) NQF #2510: All-Cause Risk-Standardized Readmission Measure," (Centers for Medicare & Medicaid Services, 2015). 5-6

**S.9.2. Risk Adjustment Type** (Select type)

Statistical risk model

If other:

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

The MSPB-PAC HH measure is stratified by Standard, LUPA, and PEP claims types. Risk adjustment is then performed separately for MSPB-PAC HH Standard, LUPA, and PEP cases. Thus, HH Standard, LUPA and PEP episodes are compared only with HH Standard, LUPA and PEP episodes, respectively, to ensure that the measure is making fair comparisons between clinically similar beneficiaries.

**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

As discussed in section S.7.2, the MSPB-PAC HH measure removes sources of variation which are not directly related to decisions to utilize care, such as local or regional price differences, to capture differences in beneficiary resource use that an HHA can influence through appropriate practices and care coordination. The MSPB-PAC HH measure relies on a detailed price standardization methodology to exclude geographic payment rate differences; in other words, the MSPB-PAC HH measure adjusts observed payments for Medicare geographic adjustment factors.[1]

Notes:

[1] QualityNet, “CMS Price (Payment) Standardization Overview” (April 2019)) https://www.qualitynet.org/inpatient/measures/payment-standardization

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

Ratio

If other:

Attachment:

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

An MSPB-PAC HH measure score of 1 indicates that an HHA had an average MSPB-PAC Amount (i.e., risk-adjusted spending level) which is equal to the national episode-weighted median MSPB-PAC Amount across all HHAs during a given performance period. An MSPB-PAC HH measure score of greater than 1 indicates that an HHA had higher average risk-adjusted spending levels compared to those of the national median HHA. For example, a measure score of 1.1 indicates that the HHA had average risk-adjusted spending levels that are 10 percent higher than the median HHA. On the other hand, an MSPB-PAC HH measure score of less than 1 indicates that an HHA had lower average risk-adjusted spending levels compared to those of the median HHA. For example, a measure score of 0.9 indicates that the HHA had average risk-adjusted spending levels that are 10 percent lower than the median HHA.

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

The detailed steps to computing the measure score are described in section S.7.2. Risk-adjustment is applied in "Step 3: Calculate Predicted Episode Payments." The purpose of risk adjustment is to compensate for patient health circumstances and demographic factors that affect resource use but are beyond the influence of the attributed provider. The MSPB-PAC HH measure risk adjustment model is adapted from the model used in the NQF-endorsed MSPB-Hospital measure, which itself is an adaptation of the standard CMS-HCC risk-adjustment model.[1,2] The MSPB-PAC HH model uses a linear regression framework and a 90-day HCC lookback period. The risk adjustment model is estimated on all MSPB-PAC HH episodes that meet the exclusion criteria.

The model is estimated separately for Standard, LUPA, and PEP episodes (see section S.7.2 for description of episode types). HH episodes are only compared to episodes of the same type (i.e., LUPA episodes are only compared to LUPA episodes, and PEP episodes to PEP episodes). This ensures that comparisons are fair, meaningful, and reflective of payment policy differences within particular HH settings.

Each provider's MSPB-PAC HH measure score is calculated as a provider's average MSPB-PAC Amount divided by the median MSPB-PAC Amount across all providers. A provider's MSPB-PAC HH Amount is defined as the sum of standardized, risk-adjusted spending across all of a provider's eligible episodes divided by the number of episodes for that provider. Below is a description of the risk adjustment variables.

Risk-Adjustment Variables

The following beneficiary health status indicators are included as covariates in each MSPB-PAC HH risk adjustment model and to the greatest extent possible are consistent across PAC settings (see Appendix C of the Measure Specifications document provided in section S.1 for a comprehensive list of independent variables used in the risk adjustment model):

- 70 HCCs
- 11 HCC interactions
- 11 brackets for age at the start of the episode
- Original entitlement to Medicare through disability
- ESRD status
- Long-term care institutionalization at start of episode.[3]
- Six clinical case-mix categories reflecting recent prior care (described further below).[4]
- Hospice utilization during the episode
- Prior acute ICU utilization day categories

- Prior acute length of stay categories

The clinical case-mix category variables used in the MSPB-PAC HH risk adjustment model are included to account for differences in intensity and type of care received by beneficiaries prior to the start of an MSPB-PAC HH episode. See section S.7.5 for more details on the methodology of assigning clinical case-mix categories to each episode.

Notes:

[1] QualityNet, "CMS Price (Payment) Standardization Overview" (April 2019)) https://www.qualitynet.org/inpatient/measures/payment-standardization

[2] CMS, "Medicare Risk Adjustment Information" (2016) https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html

[3] Identifies beneficiaries who have been institutionalized for at least 90 days in a given year. The indicator is based on 90-day assessments from the Minimum Data Set (MDS) and is calculated based on CMS' definition of institutionalized individuals.

[4] There are 7 case-mix categories as described above, but one category is removed to prevent collinearity.

**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).

MSPB-PAC HH measure scores are reported publicly for providers with 20 or more eligible episodes, along with the national average score. The distribution of MSPB-PAC HH measure scores (based on CY 2016-2017 data) that are statistically significantly different from the national average is as follows:

- Significantly lower than the national average: 34.6%
- Not statistically different from the national average: 23.4%
- Significantly higher than the national average: 42.0%

Inference about both measure performance of individual providers can be made based on the score value. The distribution of MSPB-PAC HH measure score values based on CY 2016-2017 data is as follows:

- Minimum: 0.31
- 10th Percentile: 0.78
- 25th Percentile: 0.87
- 75th Percentile: 1.05
- 90th Percentile: 1.13
- Maximum: 2.44

**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.

Each HH claim triggers its own MSPB-PAC HH episode, as described in section S.7.2. The definition of MSPB-PAC HH episodes allows episodes to overlap with hospital and other MSPB-PAC episodes. MSPB-PAC HH episodes may begin within 30 days of discharge from an inpatient hospital discharge as part of a patient's

trajectory from an acute to a PAC setting. An MSPB-PAC HH stay beginning within 30 days of discharge from an inpatient hospital will, therefore, be included once in the hospital's MSPB-Hospital measure and once in the PAC provider's MSPB-PAC measure. Aligning the MSPB-Hospital and MSPB-PAC measures in this way creates continuous accountability and aligns incentives to improve care planning and coordination across inpatient and PAC settings.

Additionally, an MSPB-PAC episode may begin during the associated services period of another MSPB-PAC episode in the 30 days post-discharge. One possible scenario occurs where, for example, an HH provider discharges a beneficiary who is then admitted by another HHA within 30 days. The HHA claim would be included once as an associated service for the attributed provider of the first MSPB-PAC HH episode and once as treatment services for the attributed provider of the second MSPB-PAC HH episode.

**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 HHAs in the United States that are Medicare-certified. Any Medicare-certified HHA that submits a PAC claim during the measure performance period can be included in this measure. The rationale for identifying this peer group is that under the Improving Medicare Post-Acute Care Transformation Act (IMPACT) of 2014, HHAs (and other PAC providers) are required to report data on quality, resource use, and other measures. The MSPB-PAC HH measure was created to fulfill the statutory requirement for HHAs to submit measures of resource use for public reporting. As such, HHAs reporting the MSPB-PAC HH measure can compare their performance relative to all other Medicare-certified HHAs in the United States.

**S.13.4. Sample size**

Detail the sample size requirements for reporting measure results.

MSPB-PAC HH measure scores are publicly reported on HH Compare for HHAs with 20 or more eligible episodes. Out of 11,427 HHAs with CY 2016-2017 episodes, 957 HHAs did not meet this minimum threshold.

**S.13.5. Define benchmarking and comparative estimates**

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

The MSPB-PAC HH measure itself is not calculated using benchmarks but rather is a comparison between a given HHA's MSPB-PAC HH Amount and national episode-weighted median MSPB-PAC HH Amount. The measure score is expressed as a ratio to that national amount, wherein a measure ratio of less than one indicates lower Medicare spending than the national median, a ratio of one indicates spending that is equivalent to the national median, and a ratio of greater than one indicates spending that is greater than the national median.

**Validity – See attached Measure Testing Submission Form**

**SA.1. Attach measure testing form**

NQF_testing_attachment_HH_Acumen_2020_04_28.docx

**Measure Number** (*if previously endorsed*)**:** 3564
**Measure Title**: Medicare Spending per Beneficiary Post-Acute Care Measure for Home Health
**Date of Submission**: 12/31/2019

**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 specificaitons (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.*

## 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: Medicare enrollment database; Minimum Data Set (MDS). | ☒ other: Medicare enrollment database; Minimum Data Set (MDS); Provider of Services File; American Community Survey; Rural Urban Continuum Codes; Home Health Compare; and Common Medicare Environment (CME) database. |

## 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 Medicare Spending per Beneficiary Post-Acute Care for Home Health (MSPB-PAC HH) measure is based on Medicare fee-for service (FFS) administrative claims and uses data in the Medicare enrollment database and Minimum Data Set (MDS). The enrollment files provide information such as date of birth, date of death, sex, reason for Medicare eligibility, and enrollment in Medicare FFS. The MDS is used to construct a risk adjustment variable, indicating beneficiaries who have been institutionalized for at least 90 days in a given year. The data elements from the Medicare FFS claims are those basic to the operation of the Medicare payment systems and include data such as date of service, date of admission, date of discharge, diagnoses, procedures, and revenue center codes. The Medicare FFS claims data files are used to identify Medicare services from home health and other settings (e.g., inpatient and outpatient hospitals) within the episode window. No data beyond what agencies submit in of the normal course of business are required to calculate this measure.

This measure submission is based on calendar year (CY) 2016-2017 data, which were the most recent data available at the time of our analyses. We used the data sources listed below to develop the claims analytic file for measure specification and testing:

- **Medicare Fee-For-Services claims and enrollment data**: We access inpatient, outpatient, carrier, skilled nursing facility, home health, durable medical equipment, and hospice claims through the Centers for Medicare & Medicaid Services (CMS) Common Working File (CWF). The data dictionary for all Medicare FFS claims, demographic, and enrollment data are available at: https://www.resdac.org/cms-data?tid%5B%5D=4931&tid_1%5B%5D=1&=Find+Data+Files. General information about the CWF is available at: https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Downloads/clm104c27.pdf.

- **Minimum Data Set (MDS)**: Acumen obtains the MDS through the Quality Improvement and Evaluation System (QIES). The data dictionary for the MDS data is available at: https://www.resdac.org/cms-data/files/mds-3.0/data-documentation.

We used two mappings to group diagnosis and procedure codes for use in identifying clinical events, implementing exclusions and applying risk adjustment:

- **Agency for Healthcare Research and Quality (AHRQ) Clinical Classifications Software (CCS) groupings for Services and Procedures**: Software is available for download at: https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp

- **CMS-Hierarchical Condition Category (HCC) mappings of ICD-9 and ICD-10 codes**: We used the Version 22 CMS-HCC mapping, which is included in the software available at: https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html.

We used five additional data sources for measure testing only, not for specification:

- **2017 American Community Survey (ACS) 5-year estimate**: We used the ACS to obtain the ZIP Code Tabulation Area (ZCTA) level measures needed to compute the Agency for Healthcare Research and Quality (AHRQ) Socioeconomic Status (SES) index score for social risk factor testing. This information is

available on the US Census website: http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

- **Rural-Urban Continuum Codes 2013**: We used this data source to construct rural-urban identifiers to test the impact of social risk factors on measure performance. These codes include county FIPS indicators which are then merged onto the episode files. Additional information on this data source can be found at: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/

- **Provider of Services Current Files (POS File)**: We used this data source to describe the characteristics of Home Health Agencies (HHAs) included in specification and testing, such as census region, ownership type, and rurality, reported in Table 1. The POS file contains data on characteristics of hospitals and other types of healthcare facilities, including the name and address of the facility and the type of Medicare services the facility provides, among other information. The data are collected through the CMS Regional Offices. General information about the POS Files is available at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Provider-of-Services/index.html

- **Home Health Compare data**: We used this data source to examine the relationship between MSPB and assessment-based quality measures. The Home Health Compare data include publicly reported HH quality measures. The data are available at https://data.medicare.gov/data/home-health-compare

- **Common Medicare Environment (CME) database**: We extracted patient-level dual eligibility information from the CME database for social risk factor testing. CMS has designated the CME database as the single, enterprise-wide authoritative source for Medicare beneficiary enrollment and demographic data. The CME database integrates and standardizes different types of beneficiary data from CMS legacy systems. The CME database receives information from the Enrollment Database (EDB) and also contains additional information not available in the EDB. Description of the CME is available at: https://www.ccwdata.org/documents/10280/19002256/medicare-enrollment-impact-of-conversion-from-edb-to-cme.pdf

## What are the dates of the data used in testing?

Calendar years 2016 and 2017

## 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 |

## 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*)

This measure is based on national data. All Home Health Agencies (HHAs) paid under Medicare's HH Prospective Payment System (PPS) and included in the HH Quality Reporting Program (QRP) were included, provided they had an eligible episode. A total of 11,427 HHAs with eligible episodes in CY 2016-2017 were included in measure specification, testing, and analysis, with differences noted in **section 1.7**. **Table 1** summarizes the frequency of HHAs by census region, ownership type, and rurality. Throughout this form, we use "K" to refer to number of providers.

**Table 3. Characteristics of HHAs included in Specification and Testing of the CY 2016-2017 MSPB-PAC HH Measure (K = 11,427*)**

| Characteristic | K (%) |
|---|---|
| **Census Region** | |
| New England | 407 (3.56%) |
| Mid Atlantic | 518 (4.53%) |
| East North Central | 2,301 (20.14%) |
| West North Central | 738 (6.46%) |
| South Atlantic | 1,783 (15.60%) |
| East South Central | 429 (3.75%) |
| West South Central | 3,013 (26.36%) |
| Mountain | 726 (6.35%) |
| Pacific | 1,462 (12.79%) |
| U.S. Territories | 49 (0.43%) |
| **Ownership type** | |
| Government | 508 (4.45%) |
| Proprietary | 9,147 (80.05%) |
| Not-For-Profit | 1,770 (15.49%) |

| Characteristic | K (%) |
|---|---|
| **Rurality** | |
| Rural | 1,825 (15.97%) |
| Urban | 9,600 (84.01%) |

Analysis of Medicare Claims File for HH CY 2016-2017 and 2016-2017 POS.

*Provider information was not available for two HHAs.

## 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*)

Measure specification and testing was based on national data. All eligible HH patient episodes with HH discharges between January 1, 2016, and December 31, 2017, were included in the measure. For patients with multiple HH episodes during the measurement period, all eligible episodes were included. A total of 10,532,450 patient episodes were included after sample exclusions were applied. Note that the MSPB score calculation removes outliers at the 1st and 99th percentile of the residual distribution and subsequent testing and analyses were conducted on these 10,321,802 episodes. **Table 2** presents demographic characteristics of the patient episodes. Throughout this form, we use "N" to refer to number of patient episodes.

**Table 4**. **Demographic Characteristics of Episodes Included in Specification and Testing of the MSPB-PAC HH Measure (N = 10,532,450*)**

| Characteristic | N (%) | Characteristic | N (%) |
|---|---|---|---|
| **Male** | 3,922,491 (37.24%) | **Race**** | |
| **Female** | 6,609,959 (62.76%) | **White** | 8,388,535 (79.64%) |
| **Male under 65 years** | 658,532 (6.25%) | **Black** | 1,410,850 (13.40%) |
| **Female under 65 years** | 752,616 (7.15%) | **Other** | 668,321 (6.35%) |

Analysis of Medicare Claims File for HH CY 2016-2017 and Medicare Enrollment database.

*MSPB-PAC HH scores were calculated after outlier episodes were removed, using 10,321,802 episodes.

** Race information is not available for 64,744 episodes (0.61% of total N).

## 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.

MSPB-PAC HH score calculation removes outliers in the 1st and 99th percentiles of the residual distribution and subsequent testing and analyses were conducted on these 10,321,802 episodes.

There were a total of 11,427 HHAs with CY 2016-2017 episodes. Agency-level performance measure score reliability and validity testing was restricted to the 10,470 HHAs with 20 or more episodes in CY 2016-2017. We applied this restriction because this measure is only to be publicly reported for providers with a minimum of 20 stays in the two-year measure calculation period.

### 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.

We analyzed the impact of the following beneficiary-level and community-level social risk factors:

- Medicare/Medicaid dual eligibility,
- Race/ethnicity,
- Urbanicity, based on beneficiary ZIP code, and
- Socioeconomic status (SES), based on beneficiary ZIP code and using data from the 2017 American Community Survey (5-year file).

---

## 2a2. RELIABILITY TESTING

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

### 2a2.1. What level of reliability testing was conducted?

(*may be one or both levels*)

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

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

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

We examined the measure score's ability to capture between-agency differences versus random error using a signal-to-noise reliability score, defined below. We also examined measure score repeatability by assessing agreement between an agency's MSPB scores based on randomly-split independent subsets of HH episodes. Performance measure score reliability testing was restricted to the 10,470 Home Health Agencies (HHAs) with 20 or more episodes in the CY 2016-2017 measurement period, as only HHAs with a minimum of 20 episodes in the two-year measurement period will be publicly reported.

**Reliability Score**: Measure reliability scores reflect the extent to which variation in the measure is due to true, underlying provider performance rather than random variation (i.e., statistical noise) due to the sample of cases observed. In the case of MSPB-PAC, the reliability score captures how much of the variance in measure scores is due to differences in episode payments between agencies rather than differences in episode payments within an agency's set of episodes. This score is calculated for each agency, using the following formula:

$$\textbf{Reliability Score: } R_k = V_b/(V_b + (V_{w_k}/n_k)),$$

where $R_k$ is the reliability for agency *k*, $V_b$ is the between-agency variance, $V_{w_k}$ is the within-agency variance for agency *k*, and $n_k$ is the number of MSPB episodes for agency *k*.[1] We report the agency-level distribution of the reliability score.

**Split-sample Reliability Testing**: This test examined agreement between two performance measure scores for an agency based on randomly-split, independent subsets of HH episodes. Good agreement indicates that the performance score is more the result of agency characteristics, like efficiency of care, rather than statistical noise due to random variation. We used four years of data (CY 2014-2017) to achieve numbers of episodes per agency in the split-half samples that are comparable to the numbers used for the actual measure scores. The sample was stratified by calendar year, thus ensuring that episodes within each calendar year were evenly distributed across the split-halves. We calculated performance measure scores for each split-half sample using the same measure specification. We then calculated Shrout-Fleiss intraclass correlation coefficients ICC(2,1)[2] between the split-half scores to measure reliability.

We also calculated ICCs between split-half scores stratified by agency size to assess whether reliability was acceptable across providers of varying sample size. To do this, we first split our sample of 10,470 HHAs into quartiles based on agency size. We then calculated ICCs within each quartile using the split-half performance measure scores derived above. Lower ICC scores indicate less correlation between the two estimates, a score of 1 would mean the estimates are exactly the same.

The Reliability Score and the ICC capture related, but distinct, concepts. ICC(2,1) will tend to differ from the Reliability Score metric for two reasons: the denominator of ICC(2,1) (i) includes statistical variation arising from true differences in a provider's performance across performance periods; and (ii) imposes a common variance for the residuals 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

---

[1] Adams J, Mehrota A, Thoman J, McGlynn E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021.

[2] Shrout, Patrick E., and Joseph L. Fleiss. "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin* 86, no. 2 (1979): 420.

half from true case sizes, mechanically reducing precision from the intended application of the measures. For these reasons, we view the Reliability Score as the preferred and more relevant metric of reliability. We still present both reliability metrics for completeness.

## 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 3** presents the mean, 25th, 50th, and 75th percentile values of the reliability scores among 10,470 HHAs from CY 2016-2017, and by sample size quartile. The average reliability score for all agencies was 0.84 and the median was 0.90. When examined by agency size, the average reliability score increased from 0.63 (quartile 1) to 0.97 (quartile 4).

**Table 5. Agency Reliability Score Distribution of the Episode-Level MSPB Risk Adjusted Spending, overall HHA sample and by sample size quartile, with public reporting exclusions (k = 10,470)**

| Agency Sample | K | Mean (SD) | 25th Pct* | Median | 75th Pct |
|---|---|---|---|---|---|
| Overall | 10,470 | 0.84 ( 0.16) | 0.79 | 0.90 | 0.95 |
| Quartile 1: 20-180 episodes | 2,623 | 0.63 ( 0.18) | 0.52 | 0.66 | 0.76 |
| Quartile 2: 181-466 episodes | 2,619 | 0.84 ( 0.06) | 0.81 | 0.85 | 0.89 |
| Quartile 3: 467-1,063 episodes | 2,612 | 0.92 ( 0.03) | 0.91 | 0.92 | 0.94 |
| Quartile 4: 1,064-68,523 episodes | 2,616 | 0.97 (0.02) | 0.96 | 0.97 | 0.98 |

* Pct = percentile. Analysis of Medicare Claims File for HH CY 2016-2017.

Note: Agency size can vary based on the sample in which the regression was estimated due to outlier exclusions.

**Split-sample Reliability Testing Results**. **Table 4** presents ICC(2,1) between the split-sample scores for the overall sample of 10,470 HHAs included in this testing, and by sample size quartile. The ICC in the overall sample was 0.76 with a 95% confidence interval (CI) of 0.75 to 0.77. The ICC was lowest in the first sample size quartile and increased progressively with increasing quartile.

**Table 6. Split-sample reliability: Intraclass correlation coefficients between split-sample performance measure scores for the overall HHA sample and by sample size quartile, with public reporting exclusions (N = 10,470)**

| Agency Sample | K | ICC(2,1) (95% CI) |
|---|---|---|
| Overall | 10,470 | 0.76 (0.75-0.77) |
| Quartile 1: 20-180 episodes | 2,623 | 0.57 (0.54-0.59) |
| Quartile 2: 181-466 episodes | 2,619 | 0.82 (0.81-0.83) |
| Quartile 3: 467-1,062 episodes | 2,610 | 0.90 (0.89-0.90) |
| Quartile 4: 1,063-68,523 episodes | 2,618 | 0.94 (0.94-0.95) |

Analysis of Medicare Claims File for HH CY 2014-2017.

Note: Agency size can vary based on the sample in which the regression was estimated due to outlier exclusions.

## 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, reliability testing results indicated good performance measure score reliability.[3] Reliability among the smallest publicly reported providers (quartile 1), is moderate to good, depending on the analysis used.

The reliability score results indicated that the average agency had good reliability. On average, 84 percent of the variation in the risk adjusted MSPB amount was associated with systematic differences between agencies[4], with a range of 63 to 97 percent (on average) among the smallest and largest agency quartiles, respectively.

The split-half reliability analysis provides further evidence of reliability and repeatability of the performance measure. Reliability (ICC) was good overall, at 0.76, with a range of 0.57 to 0.94 (on average) among the smallest and largest agency quartiles, respectively.

_____

# 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**

[3] 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). Nunnally (1978) is often cited to justify a threshold of 0.7 for "sufficient" reliability. Other authors provide other 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." 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)

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159-174.

Nunnally, J. C. (1978). Psychometric theory. New York, NY: McGraw-Hill.

U.S. Department of Education (2018), What Works Clearinghouse (WWC) Standards Handbook version 4.0.

[4] Thompson, M. P., Kaplan, C. M., Cao, Y., Bazzoli, G. J., & Waters, T. M. (2016). Reliability of 30-Day Readmission Measures Used in the Hospital Readmission Reduction Program. *Health services research*, *51*(6), 2095-2114.

☒ **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)*

To empirically test Performance Measure Score validity, we used the following methods, we:

1. Evaluated the empirical validity of the MSPB-PAC measure by examining correlation with known indicators of resource or service utilization, specifically hospital admissions and emergency room (ER) visits during the episode period. For this analysis, we compared the ratio of observed over expected spending for MSPB-PAC HH episodes with and without hospital admissions occurring in the episode period. We also compared the observed over expected spending for episodes with and without ER visits. This analysis sought to confirm the expectation that variation in service utilization is captured by the MSPB-PAC cost measure.
2. Examined the correlation between MSPB-PAC HH scores and (i) the Discharge to Community (DTC) rates and (ii) Acute Care Hospitalization (ACH) rates for CY 2016-2017. Both the Discharge to Community-Post Acute Care Measure for Home Health Agencies (DTC-PAC HHA) and the Acute Care Hospitalization During the First 60 Days of Home Health are endorsed by NQF (#3477 and #0171, respectively), are publicly reported as part of the HH Quality Reporting Program, and are based on Medicare claims.[5]

The DTC measure assesses successful discharge to community from an HHA, with successful discharge to community including no unplanned hospitalizations and no death in the 31 days following discharge. Specifically, this measure reports an HHA's risk-standardized rate of Medicare fee-for-service (FFS) patients who are discharged to the community following an HHA stay, and do not have an unplanned admission to an acute care hospital or long-term care hospital (LTCH) in the 31 days following discharge to community, and who remain alive during the 31 days following discharge to community. DTC is calculated using two consecutive years of data.

The ACH measure is, conceptually, the converse of the DTC measure, with differences in the timeframe covered. Specifically, it reports the risk-adjusted percentage of home health stays in which Medicare patients had an unplanned admission to an acute care hospital during the 60 days following the start of the home health stay. ACH is calculated using one year of data.

[5] Medicare and Medicaid Programs; CY 2017 Home Health Prospective Payment System Rate Update; Home Health Value-Based Purchasing Model; and Home Health Quality Reporting Requirements (CMS-1648-P). Available at https://www.federalregister.gov/documents/2016/11/03/2016-26290/medicare-and-medicaid-programs-cy-2017-home-health-prospective-payment-system-rate-update-home.

We hypothesized that there would be a negative association between the MSPB measure and DTC measure, indicating that providers with lower MSPB scores (more efficient providers) would have higher rates of successful discharge to the community. Conversely, we hypothesized that there would be a positive correlation between MSPB and ACH, indicating that less efficient providers have higher rates of unplanned hospital admissions. Providers whose patients have adverse events, such as re-hospitalization, at a rate higher than would be expected based on patient characteristics, should have lower DTC scores, higher ACH scores, and higher (less cost efficient) MSPB scores. The reverse should be true for providers whose patients have fewer than expected adverse events.

3. We examined the correlation between MSPB-PAC HH scores and provider's functional improvement quality measure scores publicly reported on the Home Health Compare website for CY 2017. Specifically, we use the following NQF-endorsed assessment-based measures:

    - Improvement in ambulation (#0167)
    - Improvement in bathing (#0174)
    - Improvement in bed transfer (#0175)
    - Improvement in management of oral medications (#0176)
    - Improvement in pain interfering with activity (#0177)

    These measures report risk-standardized proportions of episodes where patients' assessment scores indicate improvement between start (or resumption) of care and discharge. All are calculated using one year of data.

    The relationship between resource use and outcome measures depends on many factors, including the exact construction of each measure, payment policies, and the real-world relationship between service provision and both immediate and longer-term patient outcomes. Specifically, the relationship depends on whether better-than-expected improvement in function is associated with fewer adverse outcomes, such as unplanned hospitalizations. To answer this question, we examined the relationship between functional improvement and two other claims-based measures which capture adverse outcomes, as described earlier: DTC and ACH. We found that functional improvement is not associated with lower rates of adverse outcomes: Pearson correlations between DTC and functional improvement measures range between -0.038 and -0.003 (Spearman: -0.065 to -0.037); Pearson correlations between ACH and functional improvement measures range between -0.022 and 0.037 (Spearman: -0.049 to 0.046).[6] Accordingly, we hypothesized that the correlations between functional improvement measures and MSPB-PAC HH should, likewise, be low.

## 2b1.3. What were the statistical results from validity testing?

(*e.g., correlation; t-test*)

[6] See **Appendix Table 2b1.3** for detailed correlation results.

We found a positive relationship between MSPB and known indicators of resource or service utilization. The mean observed to expected cost ratio for episodes without a hospital admission is 0.68, compared with 2.31 for episodes with at least one hospital admission during the episode period (p-value<0.0001). The mean observed to expected cost ratio for episodes without an ER visit is 0.89, compared to 1.39 for episodes with at least one ER visits (p-value<0.0001). We also observe a positive relationship between the mean observed to expected cost ratio and the number of hospitalizations/ER visits (**Table 5**).

**Table 7. Mean Cost Ratio, by Number of Hospitalizations/ER Visits**

| Number of High-Cost Event | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Hospitalizations | 0.68 | 2.11 | 2.84 | 3.20 | 3.41 |
| ER Visits | 0.89 | 1.30 | 1.57 | 1.73 | 1.87 |

Analysis of Medicare Claims File for HH CY 2016-2017.

We also found a small, significant negative association between MSPB measure scores and the DTC measure scores (**Table 6**). Both Pearson and Spearman rank correlations revealed similar relationships.

**Table 8. Correlations between MSPB, Discharge to Community (DTC), and Acute Care Hospitalization (ACH) Measures**

| Measure Name | K* | Pearson Correlation | p-value | Spearman Correlation | p-value |
|---|---|---|---|---|---|
| Discharge to Community (DTC) | 9,711 | -0.240 | <0.001 | -0.250 | <0.001 |
| Acute Care Hospitalization (ACH) | 8,314 | 0.298 | <0.001 | 0.305 | <0.001 |

Analysis of Medicare Claims File for HH CY 2016-2017.

*Number of reflects providers with both MSPB and DTC measure scores.

Lastly, we found very small, significant positive correlations (both Pearson and Spearman) between MSPB measure scores and functional improvement measure scores (**Table 7**). These results are consistent with the fact that functional improvement is not associated with lower rates of adverse events; they may indicate that provision of additional therapy services is associated with somewhat higher rates of functional improvement, on average.

**Table 9. Correlations between MSPB and Functional Improvement Measures**

| Measure Name | K* | Pearson Correlation | p-value | Spearman Correlation | p-value |
|---|---|---|---|---|---|
| How often patients got better at walking or moving around | 8,646 | 0.128 | <0.0001 | 0.125 | <.0001 |
| How often patients got better at bathing | 8,662 | 0.163 | <.0001 | 0.150 | <.0001 |
| How often patients got better at getting in and out of bed | 8,573 | 0.153 | <.0001 | 0.152 | <.0001 |
| How often patients got better at taking their drugs correctly by mouth | 8,429 | 0.141 | <.0001 | 0.134 | <.0001 |
| How often patients had less pain when moving around | 8,572 | 0.075 | <.0001 | 0.041 | 0.0001 |

Analysis of Medicare Claims File for HH CY 2016-2017 and HH Compare data CY 2017.

*Number of reflects providers with both MSPB scores and scores for the other measure.

### 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?*)

The positive relationship between MSPB and known indicators of resource/service utilization confirms that the MSPB measure is sensitive to both the occurrence and the intensity of high cost events. The small, significant negative correlation between MSPB and DTC measures and the small, significant positive correlation between MSPB and ACH measures confirm that, on average, more efficient HHAs are associated with better discharge to community rates and fewer unplanned hospitalizations. The low correlation between MSPB and functional improvement measures is consistent with the finding that functional improvement is not associated with lower rates of adverse events (as measured by DTC and ACH measures).

---

## 2b2. EXCLUSIONS ANALYSIS

**NA ☐ no exclusions — *skip to section 2b4***

### 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*)

Measure exclusion criteria and rationale for exclusions are presented in **section 2b2.3**. We examined the episode-level frequency of each exclusion and the facility-level distribution of exclusions. The exclusions were required to ensure availability of complete and valid data for measure specification (e.g., excluding episodes in which a patient is not enrolled in Medicare FFS or where Medicare was not the primary payer for the entirety of a 90-day lookback period plus episode window).

### 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** displays the overall number and percentage of episodes excluded based on each criterion. Because the exclusions are not applied sequentially, one episode could be excluded for multiple reasons and the sum of individual exclusion frequencies may exceed the total number of episodes excluded. Overall, 19.80% of episodes were excluded because of one or more exclusion criteria. 11.60% of episodes were excluded due to a patient not being enrolled in Medicare FFS for the entirety of the 90-day lookback period plus the episode window. 7.56% of episodes were excluded because a patient had a primary payer other than Medicare for any part of the 90-day lookback period plus the episode window.

**Table 9** shows the distribution of exclusions at the facility level. On average, 12.6% of a facility's episodes were excluded due to a patient not being enrolled in Medicare FFS for the entirety of the 90 day

lookback period plus the episode window and 7.3% of a facility's episodes were excluded because a patient had a primary payer other than Medicare for any part of the 90-day lookback period plus the episode window.

**Table 10. Episode Frequencies of Exclusion Criteria for the MSPB-PAC HH Measure**

| Exclusion | N* | % |
|---|---|---|
| Any episode that results from a Request for Anticipated Payment (RAP) | 98,704 | 0.75% |
| Any episode that is triggered by a home health claim outside the 50 states, D.C., Puerto Rico, or U.S. territories | 0 | 0.00% |
| Any episode where the claim(s) constituting the attributed HHA provider's treatment have a standard allowed amount of zero or where the standard allowed amount cannot be calculated | 229,663 | 1.75% |
| Any episode in which a beneficiary is not enrolled in Medicare FFS for the entirety of the 90-day lookback period (i.e., a 90-day period prior to the episode trigger) plus episode window (including where a beneficiary dies), or is enrolled in Part C for any part of the lookback period plus episode window | 1,522,916 | 11.60% |
| Any episode in which a beneficiary has a primary payer other than Medicare for any part of the 90-day lookback period plus episode window | 992,926 | 7.56% |
| Any episode where the claim(s) constituting the attributed HHA provider's treatment include at least one related condition code indicating that it is not a prospective payment system bill | 61 | 0.00% |
| Any episode with problematic claims data (e.g., anomalous records for stays that overlap wholly or in part, or are otherwise erroneous or contradictory) | 27,951 | 0.21% |
| Total Number of Episodes Excluded | 2,600,966 | 19.80% |

Analysis of Medicare Claims File for HH, CY 2016-2017

*Exclusions are not mutually exclusive; one episode could be excluded for multiple reasons. The sum of individual exclusion frequencies may exceed the total number of episodes excluded.

**Table 11. Facility-Level Distribution of Exclusion Criteria for the MSPB-PAC HH Measure**

| Exclusion | Mean | 25th Perc. | Median | 75th Perc. |
|---|---|---|---|---|
| Any episode that results from a Request for Anticipated Payment (RAP): | 1.86% | 0.00% | 0.32% | 1.04% |
| Any episode that is triggered by a home health claim outside the 50 states, D.C., Puerto Rico, or U.S. territories | 0.00% | 0.00% | 0.00% | 0.00% |
| Any episode where the claim(s) constituting the attributed HHA provider's treatment have a standard allowed amount of zero or where the standard allowed amount cannot be calculated | 5.63% | 0.58% | 1.60% | 4.25% |

| Exclusion | Mean | 25th Perc. | Median | 75th Perc. |
|---|---|---|---|---|
| Any episode in which a beneficiary is not enrolled in Medicare FFS for the entirety of the 90-day lookback period (i.e., a 90-day period prior to the episode trigger) plus episode window (including where a beneficiary dies), or is enrolled in Part C for any part of the lookback period plus episode window | 12.65% | 7.67% | 10.38% | 14.26% |
| Any episode in which a beneficiary has a primary payer other than Medicare for any part of the 90-day lookback period plus episode window | 7.34% | 4.58% | 6.73% | 9.09% |
| Any episode where the claim(s) constituting the attributed HHA provider's treatment include at least one related condition code indicating that it is not a prospective payment system bill | 0.00% | 0.00% | 0.00% | 0.00% |
| Any episode with problematic claims data (e.g., anomalous records for stays that overlap wholly or in part, or are otherwise erroneous or contradictory) | 0.43% | 0.00% | 0.10% | 0.33% |

Analysis of Medicare Claims File for HH, CY 2016-2017

## 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*)

Exclusions for the MSPB-PAC measure are listed in **Table 10**, along with the rationale for each exclusion. Enrollment in Medicare FFS with Medicare as the primary payer is necessary to calculate the Medicare spending measures. Without these exclusion criteria it would not be possible to calculate the full cost to the Medicare program. The risk-adjustment methodology relies on Medicare FFS claims during the 90-day look back period and, therefore, the model will not be accurate without full Medicare claims data for the period of analysis.

**Table 12**. **Exclusion Criteria and Rationale for the MSPB-PAC HH Measure**

| Exclusion | Rationale |
|---|---|
| Any episode that results from a Request for Anticipated Payment (RAP). | HHA requests for anticipated payment claims are interim claims that do not reflect the final payment made by Medicare for the services. |
| Any episode that is triggered by a home health claim outside the 50 states, D.C., Puerto Rico, or U.S. territories. | This exclusion ensures that complete claims data are available for each provider. |

| Exclusion | Rationale |
|---|---|
| Any episode where the claim(s) constituting the attributed HH provider's treatment have a standard allowed amount of zero or where the standard allowed amount cannot be calculated. | Episodes where the claim(s) constituting the attributed PAC provider's treatment are zero or have unknown allowed payment do not reflect the cost to Medicare. Including these episodes in the calculation of MSPB-PAC HH measure could potentially misrepresent a providers' resource use. |
| Any episode in which a beneficiary is not enrolled in Medicare FFS for the entirety of the 90-day lookback period (i.e., a 90-day period prior to the episode trigger) plus episode window (including where a beneficiary dies), or is enrolled in Part C for any part of the lookback period plus episode window. | Episodes meeting this criteria do not have complete claims information that is needed for risk adjustment and the measure calculation, as there may be other claims (e.g., for services provided under Medicare Advantage [Part C]) that are not observable in the Medicare Part A and B claims data. Including these episodes in the MSPB-PAC measures could potentially misrepresent a provider's resource use. This exclusion also allows us to faithfully construct Hierarchical Condition Categories (HCCs) for each episode by scanning the lookback period prior to its start without missing claims. |
| Any episode in which a beneficiary has a primary payer other than Medicare for any part of the 90-day lookback period plus episode window. | Where a patient has a primary payer other than Medicare, complete claims data may not be observable. These episodes are removed to ensure that the measures are accurately calculated using complete data. |
| Any episode where the claim(s) constituting the attributed HH provider's treatment include at least one related condition code indicating that it is not a prospective payment system bill. | Claims that are not a prospective payment system bill may not report sufficient information to allow for payment standardization. |
| Any episode with problematic claims data (e.g., anomalous records for stays that overlap wholly or in part, or are otherwise erroneous or contradictory). | The episode with the most recent processing date is kept to ensure the accuracy of data elements. |

# 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 2b5.*

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with** 124 **risk factors**

☒ **Stratification by** 3 **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.

Risk adjustment is performed separately for the MSPB-PAC HH episode types listed below:

- HH Standard
- HH Low Utilization Payment Adjustment (LUPA)
- HH Partial Episode Payment (PEP)

In calculating expected spending as part of the MSPB-PAC HH measure calculation, HH episodes are only compared to episodes of the same type (i.e., HH LUPA episodes will only be compared to HH LUPA episodes, and PEP episodes to PEP episodes). This ensures that comparisons are fair, meaningful, and reflective of payment policy differences within particular HH settings. An HH Standard episode is triggered by a home health claim to which neither a LUPA nor PEP adjustment applies. An HH LUPA episode is triggered by a home health claim to which a LUPA adjustment applies, that is, when there are four or fewer visits in a home health claim. An HH PEP episode is triggered by a home health claim to which a PEP adjustment applies. A PEP is a pro-rated adjustment for shortened episodes as a result of patient discharge and readmission to the same provider within the same 60-day home health claim, or patient transfer to another HH provider with no common ownership within the same 60-day claim. If a patient is discharged to a hospital, SNF, or IRF, and readmitted to the same HHA within the 60-day claim, a PEP adjustment does not apply. A home health claim to which both a PEP and LUPA adjustment apply triggers an HH PEP episode.

The MSPB-PAC HH risk adjustment models are adapted from the model used in the NQF-endorsed hospital MSPB measure (#2158), which is itself an adaptation of the standard CMS-HCC risk adjustment model.[7,8] The MSPB-PAC HH models use a linear regression framework and a 90-day HCC lookback period. The following beneficiary health status indicators are included as covariates in each MSPB-PAC HH risk adjustment model and to the greatest extent possible are consistent across PAC settings (see Appendix C of the Measure Specifications[9] for a comprehensive list of independent variables used in the risk adjustment models):

- 70 HCCs
- 11 HCC interactions

---

[7] QualityNet, "Measure Methodology Reports: Medicare Spending Per Beneficiary (MSPB) Measure," (2015). http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1228772057350

[8] CMS, "Medicare Risk Adjustment Information" (2016) https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html

[9] https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/2016_04_06_mspb_pac_measure_specifications_for_rulemaking.pdf

- 11 brackets for age at the start of the episode
- Original entitlement to Medicare through disability
- ESRD status
- Long-term care institutionalization at start of episode[10]
- 6 clinical case mix categories reflecting recent prior care (described further below)[11]
- Hospice utilization during the episode
- Prior acute ICU utilization day categories
- Prior acute length of stay categories

Clinical case mix categories are also included in the MSPB-PAC HH risk adjustment models to account for differences in the intensity and type of care received by beneficiaries prior to the start of an MSPB-PAC HH episode. A beneficiary is assigned to a clinical case mix category using the following methodology. Taking the most recent institutional claim (by end date) in the 60 days prior to the start of an MSPB-PAC HH episode, the episode is assigned to one of the following mutually exclusive and exhaustive clinical case mix categories:

(1) **Prior Acute Surgical IP – Orthopedic** – beneficiaries who have most recently undergone orthopedic surgery in an acute inpatient hospital

(2) **Prior Acute Surgical IP – Non-Orthopedic** – beneficiaries who have most recently undergone a non-orthopedic surgery in an acute inpatient hospital

(3) **Prior Acute Medical IP with ICU** – beneficiaries who have most recently stayed in an acute inpatient hospital for non-surgical reasons and had a stay in the ICU

(4) **Prior Acute Medical IP without ICU** – beneficiaries who have most recently stayed in an acute inpatient hospital for non-surgical reasons but did not have a stay in the ICU

(5) **Prior PAC - Institutional** – beneficiaries who are continuing PAC from an institutional PAC setting (i.e., coming from an LTCH, IRF, or SNF)

(6) **Prior PAC - HHA** – beneficiaries who are continuing PAC from an HHA

(7) **Community** – all other beneficiaries

In the event that there are multiple prior claims with the same end date in the 60 days prior to the start of a PAC episode, additional logic is employed to determine the episodes' clinical case mix category. For conflicts occurring between two IP claims, the clinical case mix category corresponding to the claim with the longest length of stay (LOS) is assigned. For all other types of conflicts including those where the LOS is the same between two IP claims, the clinical case mix category is assigned using a hierarchy in the order of the categories listed above.

[10] Identifies beneficiaries who have been institutionalized for at least 90 days in a given year. The indicator is based on 90-day assessments from the Minimum Data Set (MDS) and is calculated based on CMS' definition of institutionalized individuals.

[11] There are 7 case mix categories as described below, but one category is removed to prevent collinearity.

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

Not applicable

## 2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk

(*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

### Clinical Factors

The CMS-HCC model was selected based on previous studies evaluating its appropriateness for use in risk-adjusting Medicare claims data. This model was developed specifically for use in the Medicare population, meaning that it accounts for conditions found in the Medicare population and is calibrated on Medicare Fee-for-Service (FFS) beneficiaries. In addition, the CMS-HCC model is annually updated for changes in coding practices and is exhaustive on these code sets. Because the CMS-HCC model has already been extensively tested, we focus on adapting the CMS-HCC model to the MSPB-PAC HH measure methodology.[12]

Extensive clinical review was performed by clinicians with experience providing care in HH settings, in collaboration with Medical Officers at CMS. The hospitalizations and outpatient services felt least clinically related to the HH care were excluded from resource use calculation. Services were only added to the exclusions list if there was consensus across HH and CMS clinicians.

The MSPB-PAC HH measure accounts for comorbid interactions by incorporating a number of health status interactions, such as disability and selected HCC groups (e.g., Cystic Fibrosis, Severe Hematological Disorders, Opportunistic Infections, among others). Given the fact that beneficiaries often have more than one comorbidity, the model includes commonly observed paired condition interactions, (e.g., chronic obstructive pulmonary disease [COPD] and congestive heart failure [CHF]) and commonly observed triple-interactions (e.g., diabetes mellitus, congestive heart failure, and renal failure). The beneficiary health status indicators included in the model are described in **Section 2b3.1.1**.

We conducted an investigation of Low Utilization Payment Adjustment (LUPA) and Partial Episode Payment (PEP) episodes, which represent adjustments to the standard HH payment policy and generally indicate non-standard circumstances. We tested both controlling for LUPA and PEP episode types and stratifying the sample. We found that stratification resulted in improved overall model fit and sufficient sample sizes in all three strata (standard, LUPA, and PEP episodes).

[12] 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.

We also considered controlling for patients' functional status, which helps determine HHAs' payment. However, MSPB-PAC HH is purposely designed to reflect the costs of care decisions made by HHAs, including both the intensity of home health care and later health outcomes associated with that care. Thus, adjusting for decisions made by home health agencies themselves that affect their payment, such as the functional status coded by HHA providers, would undermine the goal of the measure by masking the variation that it is intended to capture. To ensure functional status differences did not impact the validity of the MSPB score comparisons across agencies, we also estimated MSPB scores adjusted for functional status. These results revealed that controlling for functional status had a low impact on relative provider rankings (95% of providers did not change ranking quartile, with the remaining 5% changing by a single quartile).

## Social Risk Factors

A number of studies have shown that socioeconomic status is associated with the amount of resources used during the period in which patients are hospitalized, as well as during post-acute care. For example, lower-income beneficiaries are twice as likely to use home health services as Medicare beneficiaries earning higher incomes.[13] End-of-life care for Medicare beneficiaries who are Black or Hispanic is also substantially different than the end-of-life hospital services received by Medicare beneficiaries who are White. Much of the variation in end-of-life care is due to differences in utilization levels among hospitalized patients. Beneficiaries who are Black and beneficiaries who are Hispanic are significantly more likely to be admitted to the ICU than beneficiaries who are White, and minorities also receive significantly more intensive procedures, such as resuscitation and cardiac conversion, mechanical ventilation, and gastrostomy for artificial nutrition.[14]

According to a 2014 National Quality Forum report, 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.[15] 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 race.

Given the conceptual and empirical relationship between income, race, and resource use, we analyzed the impact of the following beneficiary-level and county-level social risk factors: dual eligibility, race/ethnicity, urbanicity based on beneficiary residence, and socioeconomic status (SES). We used the CMS Enrollment Database (EDB) and Common Medicare Environment (CME) to determine dual eligibility, race, and beneficiary ZIP code. Urbanicity was defined by cross-walking beneficiary residence ZIP codes to Federal Information Processing Standard Publication (FIPS) codes,[16] then cross-walking FIPS codes to Rural-Urban Continuum Codes (RUCC_2013).[17] Socioeconomic status was determined using the Agency of Healthcare Research and

[13] Kaiser Family Foundation. "Medicare Chartbook" Fourth Edition, 2010. http://www.kff.org/medicare/upload/8103.pdf

[14] Hanchate, Amresh, et al. "Racial and Ethnic Differences in End-of-Life Costs: Why do Minorities Cost More than Whites?" Archives of Internal Medicine. 2009; 169(5):493-504.

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

[16] https://www.huduser.gov/portal/datasets/usps_crosswalk.html

[17] https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/

Quality (AHRQ) SES Index[18], calculated based on beneficiary residence ZIP Code Tabulation Area (ZCTA). ZCTA was found by cross-walking the beneficiary residence ZIP code with ZCTA. We used data from the 2017 American Community Survey (5-year file) to calculate the AHRQ SES Index, with higher values indicating higher SES.

Using these data, we conducted a number of analyses for each social risk factor:

- Calculated the frequency of patients with each social risk factor;
- Calculated average Medicare spending for patients with each social risk factor;
- Assessed the difference in the measure scores estimated with and without adjustment for the social risk factors

The outcomes of these analyses are discussed 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 MSPB-PAC HH measure broadly replicates the CMS-HCC model. The literature has extensively tested the use of the HCC model as applied to Medicare claims data. It was also adopted for the NQF-endorsed MSPB-Hospital measure (#2158). Although the variables in the HCC model were chosen to predict annual cost, CMS also uses this risk adjustment model in a number of other settings (e.g., ACOs and physician QRUR programs). More information on selection of factors included in the CMS-HCC model can be found at Pope et al. (2011). During the development phase of MSPB-PAC HH, we also conducted additional analyses. Specifically, we tested controlling for LUPA and PEP episode types, instead of stratifying the sample. We found that stratification resulted in improved overall model fit and sufficient sample sizes in all three strata. We also tested stratifying by clinical case mix categories (e.g., Prior Acute Surgical IP – Orthopedic). We found that the improvement in model fit was small, while the approach also created small sample sizes in some strata. Further, we tested a longer look-back period (180 days) for identifying patients' comorbidities. This approach, resulted in lower model fit and increased number of excluded episodes. Lastly, we used bootstrapping techniques to investigate the significance of individual risk factors. We found that the risk factors that were not consistently significant across replications were clinically analogous to the risk factors that were consistently significant. As a results, we did not remove such risk factors from the model.

[18] Bonito, A. J., Bann, C., Eicheldinger, C., & Carpenter, L. (2008). Creation of new race-ethnicity codes and socioeconomic status (SES) indicators for Medicare beneficiaries. *AHRQ Publication*, (08-0029). Available at: https://archive.ahrq.gov/research/findings/final-reports/medicareindicators/medicareindicators.pdf

## 2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors

*(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

First, we examined the frequency of patients with each social risk factor (SRF). Overall, 31% of HHA episodes in the CY 2016-2017 period involved patients who were dual-eligible, 20% involved patients who were not White, and 6% involved patients in rural locations (see **Appendix Table 2b3.4b_1**). Across episode sub-samples (Standard, LUPA, and PEP), LUPA and PEP episodes are somewhat more likely to involve patients who are not dual-eligible, are White, and reside in areas with higher average SES. PEP episodes are also more likely to involve patients living in urban areas.

Second, we compared average observed per-episode spending across patients with each SRF. Overall, we found that spending is higher for patients who are dual-eligible, patients who are Black, and patients who live in urban areas. Spending is highest among patients who reside in areas with average SES in the second and third quartiles of the AHRQ SES Index (see **Appendix Table 2b3.4b_2**). Across sub-samples, treatment period spending is lowest in LUPA episodes, followed by PEP episodes, consistent with the definition of LUPA and PEP episodes.

Third, we examined risk-adjustment models with all or some of the SRFs added (see **Appendix Tables 2b3.4b_3a-e**). We found that each SRF is statistically significant, when added to the model individually as well as when added together with all other SRFs;[19] this is expected given the large sample size. However, adding SRFs, individually or together, does not substantially improve overall model fit: adjusted R-squared values increase by less than 0.01 in all cases.

Fourth, to further examine the impact of adding SRFs to the risk-adjustment model, we examined the change in individual HHA's MSPB-PAC HH measure scores when computed with and without SRFs. The results are highly correlated: correlation between baseline scores and scores adjusted for all SRFs is 0.989 (Pearson; Spearman rank correlation is 0.988) (**Appendix Figure2b3.4b_1 and Figure2b3.4b_2**). The average absolute change in provider scores is 0.02 (**Appendix Table 2b3.4b_4a**); 86% of HHA scores change by ±0.03, i.e., less than about two tenth of a standard deviation of the measure scores (**Appendix Tables 2b3.4b_4b**).

Given the minimal impact of including SRFs in the risk adjustment model on measure scores, we do not recommend adjusting the scores for these social risk factors.

## 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*)

To test the adequacy of this model, we conducted several analyses and tests:

[19] Some coefficients do not have the same sign as the absolute difference in spending between the reference group and the controlled group. For example, while spending is higher among patients who are Black, the coefficient on the Black indicator variable is negative. This is due to collinearity between the social risk factors and the clinical factors already included in the model.

1. **Model Discrimination**: We examined the model's fit (adjusted R-squared).
2. **Risk-decile testing and plots:** We calculated the distribution of episode spending by decile to examine the model's ability to predict both very low and high cost episodes. Specifically, we created a "risk score" for each episode calculated as the predicted cost values from each episode divided by the national average predicted cost value. After arranging episodes into deciles based on the risk score, we calculated the difference and ratio between predicted and observed cost for each decile.
3. **Score Distribution**: We examined the distribution of HHA-level observed and risk-adjusted episode cost.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.
***If stratified, skip to 2b3.9***

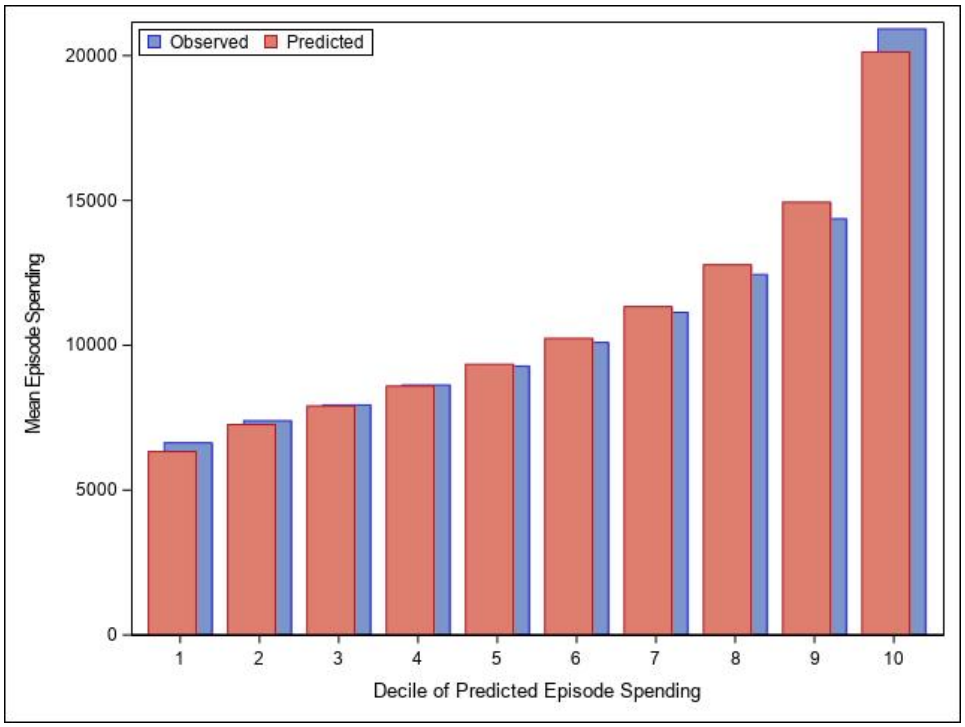
## 2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*):

The overall adjusted R-squared is 0.092. The overall adjusted R-squared was computed by fully interacting episode type (LUPA, PEP, or standard) with all other coefficients. This shows the combined explanatory power of the strata (episode types) and the covariates. The adjusted R-squared for the HH Standard episodes is 0.090; it is 0.096 for HH LUPA episodes and 0.076 for HH PEP episodes. **Appendix Table 2b3.6_1 – Table 2b3.6_3** also include regression coefficients, standard errors, and p-values for each of the covariates used in the risk adjustment model.

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

The episode-level predicted costs range from $2,562 to $69,577, indicating that this model has a range of predictions and can predict both high and low costs. The results of our analysis comparing the observed and predicted costs by risk decile are displayed in **Figure 1** and **Table 11.** In each risk decile, the observed and predicted costs are close, with a difference of 5 percentage point or less. The ratio of observed to predicted rates is close to 1 across risk deciles, with the smallest ratio being 0.96 and the largest ratio being 1.05.

**Figure 1. HH Model Diagnostics: Comparison of Observed and Predicted Spending by Predicted Spending Deciles**




Analysis of Medicare Claims File for HH CY 2016-2017.

**Table 13. HH Model Diagnostics: Comparison of Observed and Predicted Spending by Predicted Spending Deciles**

| Deciles of predicted episode cost | Number of episodes | Observed episode cost | Predicted episode cost | Predicted minus observed cost | Observed/predicted costs |
|---|---|---|---|---|---|
| 1 | 1,045,627 | 6632.12 | 6328.02 | -304.10 | 1.05 |
| 2 | 1,021,016 | 7394.15 | 7260.85 | -133.30 | 1.02 |
| 3 | 1,027,795 | 7939.56 | 7904.32 | -35.24 | 1.00 |
| 4 | 1,031,299 | 8627.22 | 8589.24 | -37.99 | 1.00 |
| 5 | 1,031,390 | 9282.35 | 9343.00 | 60.65 | 0.99 |
| 6 | 1,031,435 | 10098.20 | 10236.47 | 138.27 | 0.99 |
| 7 | 1,031,413 | 11135.70 | 11333.84 | 198.14 | 0.98 |
| 8 | 1,031,461 | 12443.52 | 12777.93 | 334.41 | 0.97 |
| 9 | 1,031,389 | 14366.97 | 14938.86 | 571.89 | 0.96 |
| 10 | 1,031,425 | 20923.21 | 20125.01 | -798.20 | 1.04 |

Analysis of Medicare Claims File for HH CY 2016-2017.

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

The results of our analysis comparing the observed and predicted spending by deciles of spending risk are displayed in **section 2b3.7** above.

## 2b3.9. Results of Risk Stratification Analysis:

Not applicable

## 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*)

The model discrimination and calibration results demonstrate good predictive ability across the full range of episodes, from low to high spending risk. There was no evidence of excessive under- or over-estimation at the extremes of episode risk. The overall adjusted R-squared is 0.092. The model controls for over 100 comorbidities (including comorbid interactions), case mix categories, and patient risk factors, and is stratified by payment-related episode types (standard/LUPA/PEP). Extensive clinical review was performed by clinicians with experience providing care in HH settings, in collaboration with Medical Officers at CMS, to identify and review relevant risk factors. Furthermore, certain features of the model improve its policy and practical usability while potentially reducing its fit statistics (adjusted R-squared). For example, controlling for potentially endogenous variables, such as therapy utilization and functional status coded by HHAs, could increase R-squared but undermine the intent of the measure by masking variation it is intended to capture (as discussed in **section 2b3.3a**). Most importantly, unrelated services, such as planned hospital admissions and routine management of certain preexisting chronic conditions (see **section S.9.1 of the Intent to Submit form**), were purposefully and carefully excluded to improve the ability to interpret and compare MSPB-PAC HH scores across providers. The R-squared cannot be evaluated alone and must be considered in combination with the costs excluded from the measure to ensure clinical validity. Since unrelated services may be well predicted by patient risk factors, excluding them can reduce the explained portion of the cost variance and the model's adjusted R-squared. For example, MSPB-PAC HH excluded services such as routine dialysis for end-stage renal disease (ESRD), as they were not felt to be prescribed by or within the scope of the home health providers. If these services had been included in the home health measure, doing so would have increased the R-squared because the ESRD indicator variable in the risk adjustment model would explain much of the variation due to dialysis. This, however, would have created an inferior measure, as it would lack clinical validity.

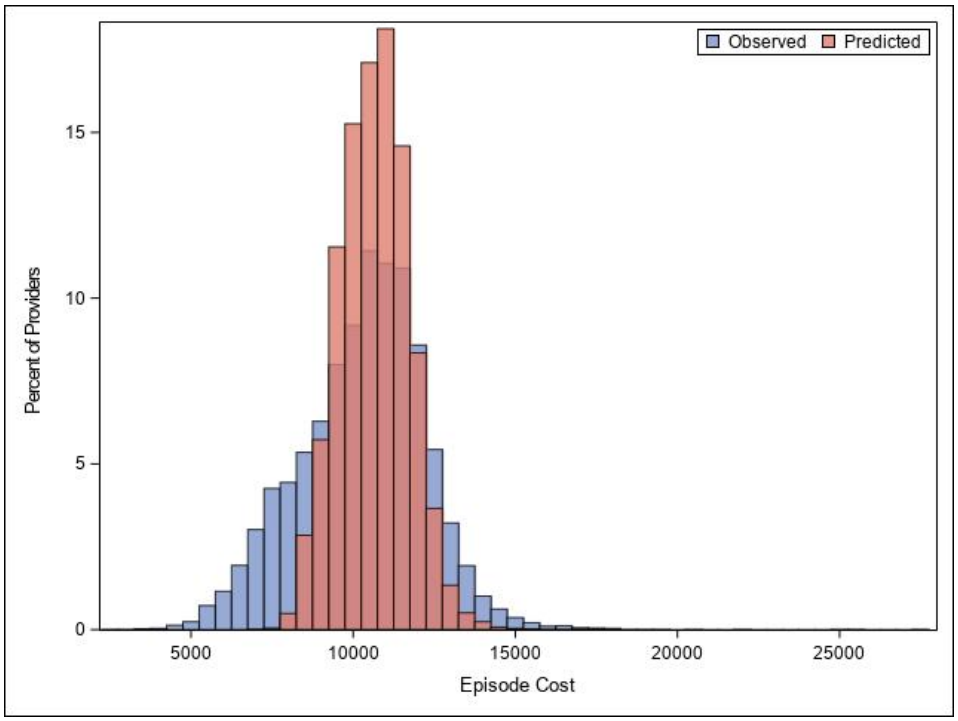
The distribution of facility-level observed and risk-adjusted spending is shown in **Table 12** and **Figure 2.** By taking into account beneficiary characteristics that are outside the provider's control, the model compresses the distribution of provider-level spending and decreases their variability. The degree of compression demonstrates that there exists a significant amount of variation in HHA spending that is not explained by the observed beneficiary risk factors.

**Table 14. Distribution of Provider-Level Observed and Risk-Adjusted Episode Spending**

| Group | K | Mean | SD | 10th Pct | 25th Pct | 50th Pct | 75th Pct | 90th Pct |
|---|---|---|---|---|---|---|---|---|
| Observed | 10,470 | 10,295.8 | 1,959.4 | 7,570.4 | 9,069.1 | 10,477.8 | 11,586.1 | 12,525.3 |
| Predicted | 10,470 | 10,648.6 | 1,052.0 | 9,295.9 | 9,896.7 | 10,666.1 | 11,371.4 | 11,949.8 |

Analysis of Medicare Claims File for HH CY 2016-2017.

**Figure 2. Distribution of Provider-Level Observed and Risk-Adjusted Episode Spending**




Analysis of Medicare Claims File for HH CY 2016-2017.

**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*)

# 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 the following two methods to identify statistically significant and meaningful differences in MSPB-PAC HH measure scores:

1. We analyzed the distribution of measure scores across all HHAs that meet the public reporting criteria. Specifically, we looked at the standard deviation of the scores and multiple points of their distribution: min, max, and 10$^{th}$, 25$^{th}$, 75$^{th}$, and 90$^{th}$ percentiles.

2. We used bootstrapping procedures to derive a confidence interval to determine if an HHA's score is significantly lower than, significantly higher than, or no different than the national average. We bootstrapped samples of HHAs, with replacement, with 1,000 replications, and re-estimated the agency average episode ratio (observed spending over expect spending) within each replication. The calculation of the expected episode spending within each bootstrap includes the addition of normally distributed noise proportional to the standard error of the agency-specific average episode spending. We use the 2.5 and 97.5 percentile of the average episode ratio from the bootstrapped distribution of each agency to calculate the full width of the 95 percent confidence interval (CI). We then compared each agency's 95% CI to the national episode-level average ratio to determine if the provider's performance was significantly different from the national mean. HHAs whose 95% CI was entirely below the national average were considered to be significantly lower than the national average; HHAs whose CI was entirely above the national average were significantly higher than the national average; and HHAs whose CI overlapped the national average were no different than the national average.

## 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 **Table 13** and **Figure 3**, MSPB-PAC HH scores are distributed fairly symmetrically and have a good deal of variability. The standard deviation is 0.15, and the max/min, 90/10, and 75/25 ratios are 7.76, 1.45, and 1.21, respectively.

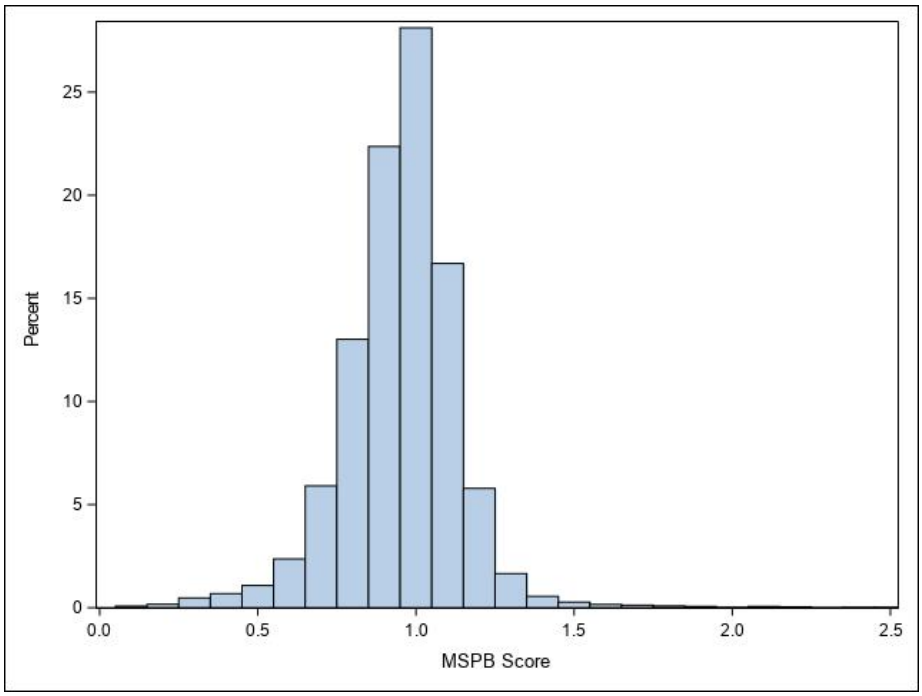
Due to the high level of reliability of the MSPB-PAC HH scores, demonstrated in **section 2a2**, small differences in scores can be interpreted as meaningful. This is confirmed by our analysis of statistical significance: 52% of HHAs had scores that were statistically significantly higher than the national mean, while 44% of HHAs had scores that were statistically significantly lower (**Table 14**).

**Table 15. Distribution of MSPB-PAC HH Scores**

| K | Mean | Standard Deviation | Min | 10th Pct | 25th Pct | 75th Pct | 90th Pct | Max |
|---|---|---|---|---|---|---|---|---|
| 10,470 | 0.96 | 0.15 | 0.31 | 0.78 | 0.87 | 1.05 | 1.13 | 2.44 |

Analysis of Medicare Claims File for HH CY 2016-2017.

**Figure 3. Distribution of MSPB-PAC HH Scores**




Analysis of Medicare Claims File for HH CY 2016-2017.

**Table 16. Proportion of HHAs with Scores Statistically Significantly Different From the National Average**

| HHA Total | Statistically significantly lower than national mean | | Not statistically significantly different from national mean | | Statistically significantly higher than national mean | |
|---|---|---|---|---|---|---|
| K | K | % | K | % | K | % |
| 10,470 | 3,618 | 34.6 | 2,452 | 23.4 | 4,400 | 42.0 |

Analysis of Medicare Claims File for HH CY 2016-2017.

## 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?*)

The MSPB-PAC HH measure is able to identify statistically significant and meaningful differences in performance across HHAs due to its good reliability and variability. Measure scores range from 0.31 to 2.44, indicating that the model can predict both low and high spending and that there are meaningful differences in agency-level spending. Seventy seven percent of HHAs have scores that are statistically significantly different from the national average, supporting the conclusion that even small difference between agency scores can be treated as meaningful.

# 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*)

Not applicable

## 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*)

Not applicable

## 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*)

Not applicable

---

# 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*)

This measure is calculated using Medicare FFS claims data; because submission and completion of claims is tied to provider reimbursement, missing data are rare. Our measure excludes episodes that are missing key measure specification data, under the exclusion criterion of claims with data that are problematic. 0.21% of episodes were excluded from our measure due to problematic claims data (e.g., anomalous records for stays that overlap wholly or in part, or are otherwise erroneous or contradictory). Thus, missing data are rare and do not have an impact on the measure. Consequently, we do not perform any formal missing data analyses.

## 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*)

As described in **section 2b6.1** above, missing data are rare and do not have an impact on the measure. Consequently, we do not perform any formal missing data analyses.

## 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*)

Not applicable

# 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.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

**F.2. Electronic Sources**

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

**F.2.1. To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*)

ALL data elements are in defined fields in a combination of electronic sources

**F.2.1a.** If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.

**F.2.2. <u>If this is an eMeasure</u>,** provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.

**Attachment:**

**F.3. Data Collection Strategy**

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

**F.3.1. Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

This measure uses Medicare Enrollment data and Medicare FFS claims from the home health, inpatient, outpatient, and physician office settings claims data, which are routinely collected for payment purposes. These data are electronically available from the Centers for Medicare & Medicaid Services (CMS) at no cost beyond that of data processing and can be used to specify, publicly report, and track the measure in a timely fashion. Since data are already collected as part of Medicare's payment process, this measure poses no additional data collection burden to providers, and because claims are used for payment, data are complete and subject to audit. In addition, this measure uses data from the Minimum Data Set (MDS) for the creation of a risk factor during risk adjustment. The MDS is necessary to construct one of the risk adjustment variables, indicating beneficiaries who have been institutionalized for at least 90 days in a given year. The submission of MDS is part of the federally mandated process for clinical assessment of all residents in Medicare and Medicaid certified nursing homes and does not pose additional burden on providers.

**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)?**

None

**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) |
|---|---|
| Quality Improvement (Internal to the specific organization) | Public Reporting<br>Home Health Quality Reporting Program<br>https://www.medicare.gov/homehealthcompare/search.html |

**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

Name of program and sponsor:
This measure is publicly reported as part of the Center of Medicare & Medicaid Services' Home Health Quality Reporting Program.
Purpose:
Section 1895(b)(3)(B)(v)(II) of the Social Security Act (SSA) requires the Secretary to establish quality reporting requirements for HHAs. More information about the HH QRP can be found at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits. In addition to tracking quality of care, quality measure data are intended to help consumers make informed decisions when selecting healthcare providers. Most quality measure data, including MSPB-PAC HH scores, from the HH QRP are publicly reported on the Home Health Compare website at https://www.medicare.gov/homehealthcompare/search.html. HH quality measure data are also available for download for providers, researchers, and other public at https://data.medicare.gov/data/home-health-compare.
Geographic area and number and percentage of accountable entities and patients included:
The HH QRP includes all HHAs paid under the HH PPS. MSPB-PAC HH scores are publicly reported for active providers with 20 or more eligible episodes in the reporting period; thus, the number of providers included in the measures can vary by reporting period. The MSPB-PAC HH measure results presented in this submission are based on 11,427 HHAs and 10,532,450 patient episodes; of these, 10,470 HHAs had 20 or more eligible episodes.

**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?*)
Not applicable – measure is publicly reported

**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.*)

Not applicable – measure is publicly reported

**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.**

Confidential feedback reports on the MSPB-PAC HH measure were provided to all active HH providers under the HH QRP starting in January 2018. These on-demand, user requested, reports are available via the internet Quality Improvement and Evaluation System (iQIES) application. Public reporting of the MSPB-PAC HH measure began in January 2019. Providers have a 30-day preview period to check their provider preview reports and submit suppression requests if there is evidence of errors in their data. CMS maintains an active provider helpdesk to which providers can submit any questions about the measure, including questions about performance data and interpretation. Individual responses are provided to each question. In addition, CMS conducts open door forums during which stakeholders can ask general questions about the measure. Along with the publicly-reported data, CMS includes consumer-friendly language to help consumers interpret measure data. Finally, MSPB-PAC HH measure specifications were publicly posted along with the CY 2017 HH PPS final rule at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/2016_04_06_mspb_pac_measure_specifications_for_rulemaking.pdf and https://www.govinfo.gov/content/pkg/FR-2016-11-03/pdf/2016-26290.pdf, respectively. The measure specifications are detailed and precise, allowing stakeholders to replicate measure calculations if they would like.

**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.**

See U.2.1.1.

Confidential feedback reports include the following data, for the provider and for the national average: reporting period, number of eligible episodes, spending during treatment period, spending during associated services period, total spending during episode, average risk-adjusted spending, national median risk-adjusted spending, and the MSPB-PAC HH score.

Publicly available data and provider preview reports include the following: reporting period, number of eligible episodes, provider MSPB-PAC HH score, and the national average MSPB-PAC HH score.

**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.**

In addition to the processes described above, we solicited public comments on the MSPB-PAC HH measure via a 24-day public comment period during January – February 2016. We posted the call for public comment on a CMS website and reached out via CMS listserv and notified TEP members. We received 45 comments during this period.

We also sought feedback on the measure through the pre-rulemaking process. We received four public comments after the release of the Measures Under Consideration (MUC) List on December 1, 2015. The MAP PAC/LTC Workgroup met on December 14-15 to consider this measure, and provided the preliminary decision of "encourage continued development" for the MSPB-PAC HH measure. Following the release of the MAP PAC Workgroup's preliminary recommendation, the report was open for a public comment period. Eight public comments on this measure were received in this time. The MAP Coordinating Committee considered these comments alongside the Workgroup recommendation and finalized the recommendation of "encourage continued development," releasing their final recommendations in February 2016. Members of the public could comment during both MAP meetings.

The measure was subject to public comment during the CY 2017 HH QRP rulemaking process. Stakeholders could comment on the specifications that were posted with the rule.

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

Comments were received from a range of stakeholders, including providers and provider associations, at each of these public comment opportunities. The comments covered a range of topics, including episode construction, exclusions, score calculation, risk adjustment, and reporting. Several commenters expressed support for the approach taken on these topics. Several commenters commented on issues such as: usefulness of setting-specific MSPB-PAC measures, usefulness of a resource use measure as a measure of quality, the adequacy of the risk adjustment model, and the process of sharing measure scores with providers. All comments were addressed, either by revising the measure or by providing the rationale why revisions are not necessary or appropriate.

The public comment summary report can be found at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/2016_03_24_mspb_pac_public_comment_summary_report.pdf.

The supplementary materials for public comment summary report can be found at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/2016_03_24_mspb_pac_public_comment_summary_report_supplementary_materials.pdf.

The MAP recommendations and summaries of public comments can be found at http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=81593

The CY 2017 HH PPS final rule with public comments and responses can be found at https://www.govinfo.gov/content/pkg/FR-2016-11-03/pdf/2016-26290.pdf.

**U.2.2.3. Summarize the feedback obtained from other users.**

Comments were received from a range of stakeholders, including researchers, government agencies, information system vendors, advocacy groups, and individuals at each of these public comment opportunities.

See U.2.2.2 for details and links to public comment summaries and the measure development team's responses to each grouping of comments.

**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**

CMS and Abt Associates team reviewed and considered all public comments during development and implementation, before finalizing the measure in the CY 2017 HH PPS final rule. Details of these considerations were provided in the public comment summary report (see link in U.2.2.2). For example, in response to public comments about the inclusion of hospice services, we added a risk adjustor for when a hospice claim begins within the beneficiary's episode window. This ensures that the HH continues to have incentives for the efficient delivery of services, but also accounts for the higher cost of episodes with hospice. We also considered public comments about risk adjusting for prior hospital stays and aligning with other IMPACT Act measures and added risk adjustors for length of prior inpatient and ICU stays.

Additionally, in response to public comments requesting more detail about the clinically unrelated excluded services, we provided detailed descriptions of the systematic process we used during development to identify clinically unrelated services.[1] This systematic process included organizing all claims into meaningful service categories, populating all services representing significant costs into a web tool used by clinicians with expertise in PAC care to determine service exclusions, and having multiple rounds of reviews to refine the list of exclusions.

We also considered other feedback that we did not implement. For example, commenters suggested controlling for community or family caregiver support. However, HH is the only setting that has community or family caregiver support information available, so inclusion of this adjustment would introduce inconsistencies between settings for the MSPB-PAC measures. Furthermore, testing indicated that the support information was more likely to be unavailable (13%) than to indicate lack of support (4% of episodes with available data). As such, community or family caregiver support did not appear to be a variable that would improve the model. For these two reasons, we did not include community or family caregiver support in the risk adjustment at this time.

[1] The process for determining clinically unrelated services is described in Appendix D of the Measure Specifications, available at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/2016_04_06_mspb_pac_measure_specifications_for_rulemaking.pdf. The complete list of excluded services is available at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits/Downloads/2016_04_06_mspb_pac_hha_service_exclusions.xlsx.

**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**

Not applicable

**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.**

Not applicable

**U.4.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**

No unexpected findings have been noted during implementation of this measure. Monitoring of patient characteristics and provider scores over time did not indicate unintended impacts on patients, to date. We are aware of the need to continuously monitor for unintended impacts on patients, such as cost-cutting at the expense of quality of care or avoiding complex patients. Our monitoring plans include monitoring trends in process and patient outcome measures, as well as trends in patient case-mix.

**U.4.2. Please explain any unexpected benefits from implementation of this measure.**

Not applicable

## Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

**H.1. Relation to Other NQF-endorsed Measures**

If there are related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

**H.1.1. List of related or competing measures (selected from NQF-endorsed measures)**

2158 : Medicare Spending Per Beneficiary (MSPB) - Hospital

**H.1.2. If related or competing measures are not NQF endorsed please indicate measure title and steward.**

**H.2. Harmonization**

**H.2.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
**Are the measure specifications completely harmonized?**
Yes

**H.2.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

**H.3. Competing Measure(s)**

**H.3.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
**Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)**
Not applicable. There are currently no measures that address both the same measure focus AND the same target population.
MSPB-PAC measures are harmonized across PAC settings as well as with MSPB-Hospital. MSPB-PAC measures were developed in parallel for all PAC settings to meet the mandate of the IMPACT Act. To align with the goals of standardized assessment across PAC settings, these measures were conceptualized uniformly across the four settings in terms of the construction logic, the approach to risk adjustment, and measure calculation. The measures mirror the general construction of MSPB-Hospital. Aligning the MSPB-Hospital and MSPB-PAC measures in this way creates continuous accountability and aligns incentives to improve care planning and coordination across inpatient and PAC settings

## Contact Information

**Co.1 Measure Steward (Intellectual Property Owner):** Centers for Medicare and Medicaid Services

**Co.2 Point of Contact:** Ronique, Evans, ronique.evans1@cms.hhs.gov, 407-786-3966-

**Co.3 Measure Developer if different from Measure Steward:** Abt Associates

**Co.4 Point of Contact:** Alrick, Edwards, alrick_edwards@abtassoc.com, 919-294-7735-

## 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.

A technical expert panel (TEP) was convened in Fall 2015. The TEP consisted of clinicians, researchers, and health care administrators with relevant expertise in PAC settings. TEP members provided input on measure

conceptualization, definitions, specifications, exclusion criteria, unintended consequences, and other considerations related to development and implementation. The TEP summary report can be found at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/Technical-Expert-Panel-on-Medicare-Spending-Per-Beneficiary_Jan-2016.pdf.

TEP members' names and organizations:

1. Alma Allen; Inova VNA Home Health, Visiting Nurse Associations of America
2. Brian Bell; Spartanburg Regional Healthcare System
3. Dexanne Clohan; Foundation for Physical Medicine and Rehabilitation Evidence-Based Practice Committee of the American Academy of Physical Medicine and Rehabilitation
4. Jean de Leon; University of Texas Southwestern Medical Center
5. Scott Guevin; Penn State Hershey Rehabilitation Hospital, AMRPA
6. Kurt Hope; Mayo Clinic, Academy of Physical Medicine and Rehabilitation
7. Steven Lichtman; Helen Hayes Hospital
8. Craig Miller; Michigan Health & Rehabilitation Services, American Physical Therapy Association
9. Mary Ousley; American Health Care Association
10. Mary Shaughnessy; Partners Continuing Care, Spaulding Rehabilitation Network and Partners Health Care at Home
11. Christopher Vaz; American Hospital Association
12. Joanne Wisely; Genesis Rehab Services, AHCA, NASL

**Measure Developer/Steward Updates and Ongoing Maintenance**

**Ad.2 Year the measure was first released:** 2018

**Ad.3 Month and Year of most recent revision:** 10, 2019

**Ad.4 What is your frequency for review/update of this measure?** Yearly

**Ad.5 When is the next scheduled review/update for this measure?** 10, 2020

**Ad.6 Copyright statement:** None

**Ad.7 Disclaimers:** None

**Ad.8 Additional Information/Comments:** Please see section Section S.8.3a for additional testing documentation.