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.

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 #: 2158

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

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

De.3. Brief Description of Measure: The MSPB Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital. Specifically, the MSPB Hospital measure assesses the cost to Medicare for Part A and Part B services performed by hospitals and other healthcare providers during an MSPB Hospital episode, which is comprised of the periods 3-days prior to, during, and 30-days following a patient's hospital stay. The MSPB Hospital measure is not condition specific and uses standardized prices when measuring costs. Beneficiary populations eligible for the MSPB Hospital calculation include Medicare beneficiaries enrolled in Medicare Parts A and B who were discharged between January 1 and December 1 in a calendar year from short-term acute hospitals paid under the Inpatient Prospective Payment System (IPPS).

IM.1.1. Developer Rationale: The MSPB Hospital measure is included in the Efficiency and Cost Reduction domain of the Hospital VBP program. With measures in other domains of clinical outcomes, safety, and person and community engagement, the HVBP program provides financial incentives to hospitals to further the value of care they provide.

The MSPB Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital. This scoring allows hospitals to improve their score by spending less than the episode-weighted risk-adjusted median cost during a given performance period through improved care coordination and provision of efficient care. For instance, hospitals can decrease (i.e., improve) their risk-adjusted episode costs through actions such as: 1) improving coordination with post-acute providers to reduce the likelihood post-discharge of adverse events, 2) identifying unnecessary or low-value post-acute services and reducing or eliminating these services, or 3) shifting post-acute care from more expensive services (e.g., skilled nursing facilities) to less expensive services (e.g., home health) in cases that would not affect patient outcomes. Care coordination helps ensure a patient's needs and preferences for care are understood, and that those needs and references are shared between providers, patients, and families as a patient moves from one healthcare setting to another. People with chronic conditions, such as diabetes and hypertension, often receive care in multiple settings from numerous providers. As a result, care coordination among different

providers is required to avoid waste, over-, under-, or misuse of prescribed medications and conflicting plans of care.

De.1. Measure Type: Cost/Resource Use

S.5. Data Source: Assessment Data

Claims

Enrollment Data

Other

S.3. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Dec 09, 2013 **Most Recent Endorsement Date:** Jul 13, 2017

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? N/A

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. High impact or high resource use:

The measure focus addresses:

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

AND

1b. Opportunity for Improvement:

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

---

1a. High Impact or high resource use.

- This measure is specified at the hospital level and evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital. It specifically assesses the cost to Medicare for Part A and Part B services performed by hospitals and other healthcare providers during an MSPB Hospital episode. Beneficiary populations eligible for the MSPB Hospital calculation include Medicare beneficiaries enrolled in Medicare Parts A and B who were discharged between January 1 and December 1 in a calendar year from short-term acute hospitals paid under the Inpatient Prospective Payment System (IPPS). This measure is not condition specific and uses standardized prices when measuring costs.
- In the previous submission in 2017, the developers demonstrated that this measure focuses on a high-priority area, the developers cited data indicating Medicare expenditures accounted for 3.6% ($647.6 billion) of the Gross Domestic Product (GDP) in 2015 and hospital benefits accounted for 30% ($188.3

billion) of those Medicare expenditures. The developer also cited data indicating Medicare expenditures will account for 6.0 to 9.1% of the GDP by 2090, if current trends continue. During the previous review cycle, the Committee agreed that the measure met the Importance to Measure and Report criterion.

- The data cited in the MedPAC Report from July 2020 shows that approximately 3,200 general short-term acute care hospitals paid under the IPPS received $189 billion in Medicare FFS revenue in 2018, increasing at average annual rate of 1.4 percent from 2014 to 2018.
- The developer notes that the scoring allows hospitals to improve their score by spending less than the episode-weighted risk-adjusted median cost during a given performance period through improved care coordination and provision of efficient care. They explained that patients with chronic conditions, such as diabetes and hypertension, often receive care in multiple settings from numerous providers. As a result, care coordination among different providers is required to avoid waste, over-, under-, or misuse of prescribed medications and conflicting plans of care.

1b. Opportunity for Improvement:

- In 2017 submission, the developer provided data from 2015 on performance trends for 3,298 inpatient prospective payment system hospitals. Measure scores ranged from 0.59 to 2.25 with an interquartile range of 0.09. These values indicate performance variation among providers.
- For the current submission, the developers provided updated data from analysis of all IPPS eligible hospitals with at least 25 episodes for the 2018 performance period, and measure score changes between 2017 and 2018.
    - The data from 2018 performance period showed a large range of provider scores on the MSPB Hospital measure with a mean of 0.99, standard deviation of 0.08, median of 0.99 and the interquartile range from 0.94 to 1.03 with the min of 0.49 and maximum of 1.68.
    - The data on measure score changes between 2017 and 2018 showed that hospital scores do vary over time; 48.8 percent of providers evidenced improved (lower) scores. The distribution in score change between these two years, with negative values indicated improvement with -1.76% and -2.01% as the 25th and 75th percentiles, respectively.

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)?

- yes
- Yes. The sheer Medicare budget (approx. 3.6% of US GDP in 2015), with hospital expenditures accounting for approximately 30%, justifies this measure which is expected to incentivize hospitals to

rein in costs. While many of the references on costs are somewhat dated, the July 2020 MedPAC Report numbers indicate that ~3200 hospitals paid under IPPS received $189 billion in Medicare FFS revenue in 2018, thus underscoring the point that the trend on hospital costs has been steadily rising over the recent years. The publicly reported MSPB measure may be keeping hospitals in check from increasing their costs even further.

- Mandated set of measures for post-acute care and there is variation in post-acute care spending
- Yes. The financial data presented in the worksheet show that this is a high impact aspect of healthcare.
- yes
- Yes
- Yes.
- Hospital spending is the largest of any sector and the develop has demonstrated it remains with very high resource use (e.g., $189 billion in Medicare FFS spending in 2018).
- Moderate

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?

- yes
- As seen from the data from 2015 and 2018, the MSPB measure did not seem to change drastically from prior submission. The IQR remained the same (0.09) in both the years, though the distribution of hospitals with improved scores in 2018 has increased from the prior year. This is somewhat expected as hospitals will shuffle in their MSPB scores year-to-year. One notable statistic is that the range of the score has gone down somewhat (0.48 to 1.69) compared to 2015 (0.59 to 2.25), potentially indicating that this measure is maturing.
- The developed found a 20% difference in 10th to 90th percentile. The measure score showed variability between 0.74 and 1.47 and that 30% of the IRFs had lower than national average and 46% of the IRFs had higher then national average.
- Yes. The measure estimates show variation - based on the range and IQR provided.
- yes
- yes
- I am not completely convinced. 1) If you look at p24-25, min and max of the 2018 score was 0.49 and 1.68 while the 10th and 90th percentile was 0.9 and 1.08. Thus the real poor performers and the real great performers are at the margins, while the majority are pretty much stacked within a very narrow range. 2) Additionally, while the developer is right to note that 48.8% had improved (lower) scores, it is also important to note by the same merit that 51.2% had worse (higher) scores.
- Episode spending is a crucial component of value-based care, whether part of ACOs or bundled payments. While this variation has decreased slightly since this measure was created, the considerable variation in episode spending that remains indicate additional opportunity for improvement.
- Moderate

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: NQF Scientific Methods Panel Subgroup **(**Evaluation A: Methods Panel**)**

Methods Panel Individual Reliability Ratings: H-7; M-0; L-0; I-0

Methods Panel Individual Validity Ratings: H-1; M-6; L-0; I-0

- This measure was reviewed by the NQF Scientific Methods Panel for reliability and validity. The NQF SMP subgroup accepted the preliminary analysis decisions for measure #2158 without further discussion. This measure passed with high rating on reliability and moderate rating for validity.

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

Evaluators: N/A

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 MSPB Hospital measure uses Medicare Part A and Part B claims data maintained by CMS.
- The developer used MSPB Hospital episodes from performance period 2018. The developer included episodes from performance period 2017 for select cross-year reliability testing.
- Reliability testing was conducted at the measure score-level:
    - The developer conducted signal-to-noise and multi-sample (or split-sample) analyses to assess reliability of the measure
        - The developer reported a mean reliability score for hospitals with at least 25 episodes of 0.92
        - The median reliability score for hospitals with at least 25 episodes was 0.96 and the reliability score interquartile range spanned from 0.91 to 0.98
        - The Pearson correlation coefficient was 0.83 for the 2018 split-sample and 0.79 for the 2017 and 2018 sample. The Shrout-Fleiss intraclass correlation coefficients were similar at 0.83 and 0.79 for the 2018 split-sample and 2017 and 2018 sample

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?*

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:

- this measure can only be implemented by the developer unless they release code for others to use
- None
- Claims based measure
- None.
- n/a
- It is not clear to me how considerations for specialty hospitals that have a disproportionally large patient volume requiring high cost care (cancer treatment, orthopedics- joint replacement) compared to a general acute hospital with a greater mix of low and high cost admissions are made.
- The measure being limited to episodes greater than 25 has made the reliability better. The availability of the code to the public is wise; however I am not sure how many institutions can implement that.
- 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.

- I agree with the SMP, nothing materially new
- None.
- The developers used signal to noise analysis and reported mean reliability score of 0.86 and median of 0.89. They showed that 86% of the variation in the risk adjusted MSPB amount was associated with systematic differences between facilities. They showed a range of 70 to 96% among the smallest and largest facility quartiles
- None. Testing results provided look satisfactory.
- yes

- No additional comments.
- I am satisfied with the S/N and split sample analyses done and their results.
- 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. Disparities: If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender), OR rationale rationale/data justifies why stratification is not necessary or not feasible.

2b1. Specifications Align with Measure Intent:

- Attribution:
    - This measure is attributed to hospitals. This attribution approach was developed in order to encourage hospitals to facilitate care coordination and support their role in reducing unnecessary resource use and costs during the period immediately prior to, during, and in the 30 days after hospital discharge.
- Costing approach:
    - The MSPB Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital.

- 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:

- *Expert Panel:* The developer examined potential refinements to the measure via an 20-member expert panel and a review of public comments.
  - The developer noted that "no official vote was taken", but panelists "agreed" with the measure's "all-cost approach", having readmissions to trigger new MSPB Hospital episode, and updating the measure's numerator amount calculation.
  - The developer further stated that panelists also provided additional considerations for ongoing social risk factor testing.
- *Empirical Validity:* The developer undertook three approaches to empirical testing
  - The developer compared costs of episodes with and without post-admission events (e.g., post-acute services) expected to increase cost.
    - The developer demonstrated that episodes with downstream readmissions, post-acute costs, and post-acute SNF costs had higher observed/expected ratios. These empirical results are consistent with the hypothesized direction.
  - The developer examined the relationship between a hospital's average expected episode cost and average episode rates of several service use categories.
    - The developer reported that the correlations across all services categories average 0.487 with procedure use having the strongest correlation of 0.721.
  - Lastly, the developer examined the relationship between the measure and other cost and efficiency-specific measures and measures in other hospital value-based purchasing (HVBP) program domains.
    - All three measures capturing 30-day Medicare payments for acute myocardial infarction (AMI), heart failure (HF), and pneumonia (PN) conditions, were positively but weakly correlated with the hospital average predicted episode cost.
    - For the Timely and Effective Care measures, which capture the time spent in the emergency department before being sent home or admitted, were also positively and weakly correlated with average predicted episode costs.
    - Measure within the HVBP Safety domain measures were positively and weakly correlated with HVBP Clinical Outcome survival rate measures, and negatively and weakly correlated with HCAHPS survey questions on hospital staff communications, cleanliness, and care transition planning.

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

- The developer excludes:
  - Transfer- or death- related episodes
  - Non-IPPS, Non-acute, or Critical access hospitals
  - Inpatient facilities in excluded states and territories
  - Episodes invalid or incomplete data
- Roughly 37% of all episodes were excluded, with the largest contributor being episodes where the initial inpatient stay was in a non-acute hospital or a critical access hospital (11.45%).

2b4/2c. Risk adjustment

- The developer controlled for case mix using a statistical risk model with 109 risk factors
- The risk adjustment model followed the CMS Hierarchical Condition Category (HCC) risk adjustment methodology used in Medicare Advantage, including 79 HCC risk factors derived from claims 90 days prior to episode start date.
- The developer used data from or based on the American Community Survey (ACS), and Common Medicare Environment (CME) in evaluating patient cohort and social risk factors in risk adjustment.
    - The developer analyzed race, sex, dual status, income, education, unemployment, the AHRQ SES Index, and the Area Deprivation Index (ADI).
    - The developer reported that models that include the AHRQ SES Index and models that include the ADI have p-values less than or equal to 0.05 in at most 10 of the 26 major diagnostic category stratifications.
    - The developer performed a decomposition analysis to assess the effects of select social risk factors between hospitals and beneficiaries.
    - The developer did not include social risk factors in the model, reporting that including social risk factors in risk-adjustment model would mask provider differences based on the decomposition analysis conducted. The developer also reports a minimal impact on measure scores from social risk factors.
- The developer reports a range of R-squared values for the measure's risk models from 0.11 to 0.67 with an overall R-squared of 0.457 and an overall adjusted R-squared of 0.456.

2b5: Meaningful Differences

- The developer reports a distribution of measure scores showing that the 90th percentile is over 21% greater than the 10th percentile with differences in rural vs. urban areas and teaching hospitals vs. non-teach hospitals.

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., correlations, exclusions, risk-adjustment approach, etc.)?*
- *Does the SC have any concerns related to the risk adjustment model (e.g., the r-squared values, lack of social risk factor adjustment)*

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

Committee Pre-evaluation Comments:
Criteria 2b: Validity

2b1. Additional threats to validity: 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?

- it is a very complicated measure with lots of formulas. The main issue is that it can take you away from reality. However, using measure development standards they are in line with the pack
- None beyond what was indicated by some of the members in the Methods Panel (#1, #2, and #5).

- Concern about social risk factors and if providers had large numbers of social risk factor patients would this negatively impact them
- None.
- na
- It is clearly described how outlier cases are excluded or accounted for. However, I am not sure how specialty hospitals who may have a disproportionately large number of high-cost episodes are accounted for. These specialty hospitals (outlier hospitals) with higher cost episodes may have an ability to reduce episode cost to a greater degree than non-specialty hospitals (non-outlier hospitals) and it may warrant looking at/measuring outlier hospitals differently to account for this along with the outlier cases that may not be so apparent if all episodes are more expensive.
- No major issues with attribution, costing or truncation. Although this measure is different in its attribution for "readmissions" called as "re-hospitalization" in the submission. This difference may be an issue to some.
- The attribution methodology accounts for potential issues that might limit an accountable entity from having reasonable control over costs measured (e.g., excluding hospital to hospital transfers). Costing approach and truncation are appropriate.
- No concerns

2b2. 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?

- The testing is standard for these measures. The measure correlations are not that helpful. What is the hypothesis around the correlations? should talk about how they view face validity. they do face validity for inputs but we need it for the outputs
- None.
- The developers found the mean observed to expected cost ratio for episodes without a hospital admission to be 0.91 and 1.39 for episodes with at least one hospital admission during the episode period.
- None.
- no concerns
- yes
- I am not satisfied with the insouciant approach to face validity. "No official vote was taken", "panelist agreed". The developer should be more rigorous. 2) I am satisfied with the empirical testing approach.
- Yes, 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:

- none
- None

- No concerns
- None. Exclusions look reasonable.
- no concerns
- I do not have any concerns with the exclusion criteria
- No major concerns with exclusions.
- While the measure as currently constructed appropriately excludes deaths that occur during the inpatient stay or shortly after discharge, I wonder if there is a way in the future to account for these episodes as well (given the often high spending that occurs in the final weeks of life).
- No concerns

2b4/2c. 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?

- This one is a difficult one. The choice not to include SES will be worth discussing. With partial effects in model this big it is hard to determine if it would have been helpful. SES matters for access to post-acute care.
- Yes.
- As noted earlier, social risk factors
- Yes, there is a conceptual relationship. I generally agree with the risk adjustment methodology. I initially had a concern about excluding all SRFs. But it appears that they didn't have a significant impact on the scores.
- acceptable
- Risk adjustment is appropriate with acceptable results.
- As with prior measures, social factors were predictive of MSPB; however were not included in the model due to inconsistent directionality of their associations. There may be several reasons for that including the lack of appropriate fit of the model as reflected in the low R2 and a modicum of granular data. However due to the inconsistencies of the associations, I am not majorly opposed to not including them.
- While I initially had concerns about using only a 90-day look back of claims for HCC risk adjustment calculation, the measure worksheet adequately addresses this issue.
- No concerns

2b5. Threats to Validity: Meaningful Differences Describe any concerns with the analyses demonstrating meaningful differences among accountable units:

- no concerns
- None
- No concerns
- No concerns.
- no concerns

- No concerns
- This is where I think the measure struggles a bit. I talked about the opportunity for improvement as above. Added to that are the weak associations of the empirical testing. While the O/E values have consistent directionality, they are rather weak. As a result, I think that like this measure will find meaningful differences in only the tail ends of the distribution which is where the greatest gains/losses will happen.
- It is encouraging that differences among accountable entities have decreased over time (e.g., the difference between 90th and 10th percentiles has trended down) and provides some support that the measure is both useful and utilized, a large enough difference remains (21%) that the measure remains valid.
- No concerns

2b6. 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 population still be valid?

- none
- None
- SES and low R2 may suggest that other factors that are not accounted for are driving episode spending
- No concerns.
- no concerns
- Missing data is accounted for leads to appropriate exclusion of episodes for the measure
- No major concerns.
- No concerns
- No concerns

Criterion 3. Feasibility

3. Feasibility

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

- The developer states that data are generated by and used by healthcare personnel during the provision of care and are coded by someone other than person obtaining original information.
- The developer states that all data elements are in defined fields in a combination of electronic data sources.
- The developer notes that CMS uses Medicare administrative claims data that hospitals submit to CMS for payment to calculate the MSPB Hospital measure. As a result, the required data are readily available and retrievable without undue burden.

- The developer indicates that there are no fees, licensing, or other requirements associated with this measure.
- In 2017 measure review, one of the Standing Committee members raised a concern that while the measure is feasible for entities like the Centers for Medicare and Medicaid Services, it would be difficult for other entities to calculate the measure independently.
- In this submission, the developer notes that the SAS code and documentation for the measure calculation is publicly available.

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):**
   - The main issue in usability is the complexity of the algorithms. The only way to allow others to calculate would be to share code.
   - None
   - Claims based measure
   - No concerns. The data elements are easily available, from electronic sources.
   - no concerns
   - Clinical outcomes are not routinely reported and remain challenging to interpret using billing data. Collection of and reporting of patient reported outcomes (PROs) has been challenging for many reasons yet is an important measure of quality. Historically PROs are not regularly collected prior to an episode (baseline score) or following the admission unless part of a research study. Additionally, many PRO scores are not expected to show improvement until after to 30-day episode ends. Standardization of PROs using a general health questionnaire such as the PROMIS 10 may be beneficial in helping us to better compare the quality and cost (Value) of the admission but barriers such as collection methods, PRO type, data storage and submission exist.
   - No major concerns.
   - Related to a concern raised by a Scientific Methods Panel member in 2017 regarding use in non-Medicare populations, outside of Medicare this measure largely becomes a utilization, and not cost metric, since prices vary widely beyond the limited variables that exist in Medicare (wage index, GME, DSH, etc.). Such a utilization metric may still be useful in commercially insured populations but can't as easily be assigned a dollar figure and be compared as such from hospital to hospital.
   - No concerns

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 included in the Efficiency and Cost Reduction domain and used within the Hospital Value-Based Purchasing (HVBP) Program. The HVBP program provides financial incentives to hospitals to further the value of care they provide based on their performance on selected quality measures.
- The developer noted that the MSPB Hospital measure is reported publicly on CMS' Hospital Compare website.

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

- The providers are given Hospital-Specific Reports (HSR) once a year, in early to mid-summer (but they can request re-uploads of their HSRs as needed) that contain information on the MSPB Hospital measures and provides information on the hospital's performance on the MSPB Hospital measure and cost breakouts of measure components in relation to state and national statistics. It was noted that CMS provides an annual webinar during which the MSPB Hospital measure methodology and measure score interpretation is detailed.
- The developer stated that the providers and other stakeholders have an opportunity to provide feedback through question & answer sessions as part of the annual webinar. CMS also provides email help-desk support for operations and other questions. Feedback is centered around methodological questions off clarifications.
- The developers noted that the potential refinements to the MSPB Hospital measure methodology that is in current use were identified from prior rule comments, past NQF endorsement cycles, and related measure development (e.g., MSPB Clinician). These potential refinements were tested and reviewed

by a Technical Expert Panel (TEP), comprised of 20 members, in February 2020 as part of the internal MSPB Hospital measure's re-evaluation.

- In 2017 review of this measure, the committee members raised concerns that the reports provided on this measure may not be fully actionable, as the information provided does not provide adequate details to show where improvement efforts should be focused. The Committee suggested the measure's usability could be enhanced by providing a more detailed breakdown of utilization by major diagnostic categories in the measure summary reports that are sent to providers.
- The developer has noted that the Hospital-specific reports do contain breakdowns of cost by major diagnostic categories and further provide patient-level data for deeper analysis.

Additional Feedback:

- N/A

***Questions for the Committee**:*

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

Staff preliminary rating for Use: ☒ Pass ☐ No Pass

Usability

4b. Usability.

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

4b.1 Improvement.

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

4b2. Benefits vs. harms.

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

4b1. Improvement results

- The developers noted that when comparing MSPB Hospital measure scores between 2017 and 2018, the data demonstrated that nearly half, 48.8%, of all hospitals improved on their MSPB Hospital measure score. The developers interpreted that the MSPB Hospital measure is able to effectively capture provider risk-adjusted spending during an episode and is able to capture differences between providers.

4b2. Unintended consequences

- The developers explained that no unintended consequences to individuals or populations have been identified during testing, and no evidence of unintended negative consequences to individuals or populations have been reported since implementation.

4b2.Potential harms

- No potential harms were identified

***Questions for the Committee**:*

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

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

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

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

- Yes, it is used
- This is an ongoing measure, and is used in HVBP.
- No concerns
- The measure is reported publicly on the Hospital Compare section within CMS's website; and used in the used within the Hospital Value-Based Purchasing program.
- no concerns
- There appears to be alignment between hospital and clinician measures. Similar data is used in CMS bundled payment programs such as CJR and BPCI-A, however those programs look at a 90-day episode and are not reported in the same format publicly. Better alignment with the measurement and reporting of these programs would reduce confusion for providers when developing program goals, would allow providers to develop internal programs with more focus on providing improved value and not have to deal with different measures for different programs.
- Yes.
- The measure continues to be used in multiple high-profile areas (HVBP program and Hospital Compare).
- Currently used in public accountability programs

4a2. Use - Feedback on the measure: 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?

- Feedback appears to be included
- The feedback on the measure by the providers has been adequately handled and incorporated by the measure developer (described in U2.2.1-U2.3.
- Not sure if and how the IRFs are using this to improve care
- It appears that all stakeholders have an opportunity to provide feedback through question & answer sessions as part of an annual webinar organized by CMS. Refinements in measure methodology are made based on the feedback received.

- no concerns
- Data sharing and feedback is appropriate
- No major concerns.
- No concerns.
- No concerns

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

- This is an area where I would like more information. A hospital level measure with so many exclusions and adjustments would be hard to use for improvement. Apparently the developer gives hospitals breakdowns of where they are higher - it would be nice to see a sample report. Just because costs have lowered does not mean that this measure contributed to that outcome. In this area, a survey of users might have been more appropriate than an empirical analysis. NQF should think about how to include USER experience in the USABILITY evaluation for maintenance measure.
- Yes.
- No improvement data was provided. Not sure that public reporting has impacted IRF costs
- Yes, About half of all hospitals improved on the MSPB measure score between 2017 and 2018.
- no concerns
- See my previous notes re: PROs. Data reported is based on readmissions really being the only indication of a poor outcome additionally measuring/comparing longer term outcomes using PROs will help to drive Value in care.
- I am not sure if this can drive improvement due to my observations above.
- Use of the measure, including measurable results over time, have been observed. I wonder if there would be benefit from more granular reporting on some aspects of episodes, specifically post-acute care spending. This might help hospitals and policy makers better understand some of the key drivers of the high variation in spending among accountable entities, as well as potential opportunities for improvement.
- Yes

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

- none
- None
- Not seen yet but the IRfs with high patients with social risk factors could show negative incentives to admit these type of patients
- None.
- no concerns
- NA
- No major concerns.

- No concerns.
- No concerns

Criterion 5: Related and Competing Measures

- The developers identified the following related and competing measures:
    - 3574: Medicare Spending Per Beneficiary (MSPB) Clinician (this measure is no longer NQF-endorsed)
    - Medicare Spending Per Beneficiary (MSPB) PAC measures:
    - 3561: Medicare Spending Per Beneficiary – Post Acute Care Measure for Inpatient Rehabilitation Facilities
    - 3562: Medicare Spending Per Beneficiary – Post Acute Care Measure for Long-Term Care Hospitals
    - 3563: Medicare Spending Per Beneficiary – Post Acute Care Measure for Skilled-Nursing Facilities (this measure is no longer NQF-endorsed)
    - 3564: Medicare Spending Per Beneficiary – Post Acute Care Measure for Home Health Agencies (this measure is no long NQF-endorsed)

Harmonization

- The developers noted that the measure specifications have been harmonized to the extent possible with the related and competing measures.
- Furthermore they note that the MSPB Hospital measure has been harmonized with MSPB Clinician and MSPB-PAC in the following ways: (i) change in risk adjusted ratio calculation, and (ii) allowing readmissions to trigger an episode (specific to MSPB Clinician).
- They stated that the MSPB Hospital measure differs from MSPB Clinician and MSPB-PAC in that it captures all Medicare Part A and Part B costs associated with an episode that is triggered by an inpatient stay while MSPB Clinician, for example, excludes services that are unrelated to clinician care.

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?

- no concerns at this time
- Yes, and the developer has explained the steps taken to harmonize the MSPB measure with other relevant measures.
- No competing measures identified
- Yes, there are related and competing measures. The developer appears to have harmonized all measure specifications as much as possible.
- no concerns
- See previous note re: MSPB vs. Bundled payment programs
- No major concerns.
- No concerns.
- None

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/26/2021

- Comment by: American Medical Association

  The American Medical Association (AMA) requests that the Standing Committee discuss the revisions made to the measure as described in S.7.2, specifically the change to equally weigh all risk-adjusted hospital episodes by the average ratio of observed to expected costs, and the expansion of episodes to include re-hospitalizations within 30 days of discharge of any admission that opens an episode. No rationale was provided for any of these changes, which makes it difficult for the AMA to provide input and determine whether we agree with the changes. The AMA is particularly concerned that the expansion to include re-hospitalizations will now double count the costs attributed to a hospital.
  The AMA does not believe that the current risk adjustment model is adequate due to the unadjusted and adjusted R-squared results ranging from 0.11 to 0.67 across the Major Diagnostic Category. The measure is not adequately tested and adjusted for social risk factors. It is unclear why the measure developer would test social risk factors after adjusting for clinical risk factors rather than assessing the impact of both clinical and social risk factors in the model at the same time. These variations in how risk adjustment factors are examined could also impact how each variable (clinical or social) perform in the model and remain unanswered questions. In addition, we note that hospitals measure scores shift when some or all of the social risk factors are applied within the risk model and particularly just over 15% of safety-net hospitals move above or below the delta in Model 13 (Table 2b34b.c Impact of Social Risk Factors). We ask the Standing Committee to carefully consider whether these results impact the ability of the measure to meet the validity criterion.
  Lastly, we would like to express our appreciation that the measure developer completed correlations with existing hospital quality measures and encourage the measure developer to continue to provide this information for other cost measures.

- No NQF Members have submitted support/non-support choices as of this date.

Combined Scientific Methods Panel Preliminary Analysis of Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form
Measure Number: 2158
Measure Title: Medicare Spending per Beneficiary (MSPB) Hospital

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**
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.)

Panel Member #2: Although previously endorsed, this version has some important changes as described in S.13.1: *"… version under consideration allows acute care hospital readmissions to trigger a new MSPB episode and changes the calculation of the MSPB Amount (the measure score numerator) from a calculation based on the ratio of average observed episode cost to average expected episode cost to an average ratio of observed to expected episode cost (see the end of Section S.7.1 for more detail)"*

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.*

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

Panel Member #2: None.

Panel Member #3: No concerns – this measure is widely used to assess total cost of inpatient care.

Panel Member #4: None

Panel Member #5: Why is the performance period "discharges between January 1 and December 1"? Is this a typo, i.e., December 31? If not, the explicit rationale for omitting December discharges needs to be provided.

Panel Member #6: 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 hospital spending. It likely has applicability only to the CMS hospital compare program for public reporting.

Panel Member #8: No concerns. Exclusion criteria is standard for Medicare measures.

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

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

Submission document: Testing attachment, section 2a2.2

Panel Member #1: Split sample analysis, using Pearson, Spearman and shrout-Fleiss ICC statistics, comparing split samples from 2018 and 2017 to 2018

S/N analysis

Panel Member #2: Signal-to-noise and multi-sample (or split-sample) analyses were conducted to assess reliability of the measure.

Panel Member #3: Signal-to-noise and multiple sample analysis – both appropriate. The random sample and ICC for the 'confirmation' sample is good (method 2), but I would also like to see a split of hospitals as well. In other words, take a random sample of hospitals and compare to the 'test' hospitals, preferably over multiple years.

Panel Member #4: Developer used the formula for signal-to-noise reliability and multi-sample testing

Panel Member #6: The developers used 2 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). 2) multi-sample reliability testing to examine agreement between 2 scores for a facility based on randomly-split, independent subsets of hospital episodes in the 2018 measurement period, and between scores for the 2017 and 2018 samples. Good agreement indicates the performance score is more the result of facility characteristics (efficient care) than statistical noise due to random variation. Only providers meeting an episode minimum of 25 episodes were included. They analyzed score agreement from Pearson, Spearman, and Shrout-Fleiss intraclass correlation coefficients ICC(2,1), where coefficients close to 1 indicate high agreement between samples.

Panel Member #8: Signal-to-noise using within-hospital variance, between-hospital variance, and the ratio of between-group variance and within-group variance. A ratio closer was interpreted as representing the impact of systematic differences between hospitals.

Also, randomly split set of episodes from 2018 performance period and the 2018 and 2017 performance periods was calculated for those providers meeting a minimum episode of 25. Pearson, Spearman, and Shrout-Fleiss intraclass correlations were measured.

Panel Member #9: Signal to noise and split sample correlation analysis was appropriate.

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

Submission document: Testing attachment, section 2a2.3

Panel Member #1: 2018 Split Pearson 0.8265, Shrout-Fleiss 8.264

2017 to 2018 Split Pearson 0.791, Shrout-Fleiss 0.7873

S/N Mean 0.92, Median 0.96, 25th Pct 0.91, 75th 0.98, all for hospitals with at least 25 episodes

Panel Member #2: The median reliability score for hospitals with at least 25 episodes was 0.96 and the reliability score interquartile range spanned from 0.91 to 0.98

The Pearson correlation coefficient was 0.83 for the 2018 split-sample and 0.79 for the 2017 and 2018 sample. The Shrout-Fleiss intraclass correlation coefficients were similar at 0.83 and 0.79 for the 2018 split-sample and 2017 and 2018 sample.

Both of these two sets of reliability exercise indicate high reliability of the measure.

Panel Member #3: Strong reliability score; good ICC correlation statistics.

Panel Member #6: The average reliability score of hospitals with at least 25 episodes was 0.92, with 94.3 percent of providers meeting or exceeding a 0.7 reliability score. While higher episode-minimums yield higher reliability results, the application of higher episode-minimums reduces the number of providers receiving a measure score. The median reliability score for hospitals with at least 25 episodes was 0.96 and the reliability score interquartile range spanned from 0.91 to 0.98. The Pearson correlation coefficient was 0.83 for the 2018 split-sample and 0.79 for the 2017 and 2018 sample. The Shrout-Fleiss intraclass correlation coefficients were similar at 0.83 and 0.79 for the 2018 split-sample and 2017 and 2018 sample. Overall, the reliability of the MSPB Hospital measure is high when its current 25-episode minimum is applied to balance measure reliability and inclusiveness. The correlation coefficients for scores across 2 years were lower than scores in the randomly split 1 year sample but this would be expected and reliability is still high across years for same hospital.

Panel Member #8: The reliability scores for providers with at least 25 hospitals was a mean of 0.92, 25th percentile of 0.91 and 75th percentile of 0.98. The split-sample Pearson correlation coefficient was 0.83 for the 2018 split sample and 0.79 for the 2017 and 2018 sample. The Shrout-Fleiss correlations coefficients were similar.

Test-retest of the 2014 and 2015 data demonstrated that low performers (40th percentile) in those timeframes were the same low performers in the other sample.

Panel Member #9: All testing supported reliability, I agree with Measure Stewards assessment

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

Submission document: Testing attachment, section 2a2.2

☒ **Yes**

☐ **No**

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

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

Submission document: Testing attachment, section 2a2.2

☒ **Yes**

☐ **No**

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

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

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

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

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

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

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

Panel Member #1: Split sample and S/N tests all high and well above minimum standard for acceptability.

Panel Member #2: As indicated in #7 above, the results from reliability tests indicate high reliability of the measure.

Panel Member #3: Good reliability; I'm giving the measure a high b/c of O/E analysis and the ICC scores – they were above 0.70 even as you go further out in time. Other measures also strong.

Panel Member #4: This submission demonstrates integrity in the determination of episode minimums for high reliability.

Panel Member #5: Reliability testing was adequate. Year-to-year distributions by decile result in comparable values. Signal-to-noise and split sample reliability values were strong.

Panel Member #6: Based on reliability testing using different approaches, all show high reliability with facility sample of at least 25 episodes.

Panel Member #8: The reliability of the measure score is in the 80% range for test-retest low performers and above 90% for those with greater than 25 episodes in the recent sample.

Panel Member #9: No concerns.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

Submission document: Testing attachment, section 2b2.

Panel Member #1: Exclusions appear conceptually reasonable but substantial number of exclusions due to death while IP or immediately after (16%) and 11.5% in non-Acute or critical access hospitals. Patients who died within the hospital had substantially higher mean costs and large standard deviation in costs than average over full sample or included cases.

Panel Member #2: None

Panel Member #3: I understand that Critical Access Hospitals face different reimbursement rules as compared to short-terms acute care hospitals. However, I would consider keeping them in the measure of having a CAH version.

Panel Member #4: None.

Panel Member #6: NONE

Panel Member #8: The exclusions are clearly defined. About 37% of all episodes were excluded, with the largest contributor being episodes where the initial inpatient stay was in a non-acute hospital or a critical access hospital.

Panel Member #9: No concerns

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

Submission document: Testing attachment, section 2b4.
Panel Member #1: None. Observed to expected ratio varies from 0.51 at 10th percentile and .63 at 25th percentile to 1.2 at 75th percentile and 1.72 at 90th. This is a wide range in the O/E rations
Panel Member #2: None
Panel Member #3: None
Panel Member #4: None
Panel Member #5: Based on the 2018 performance (see MIF form), there is a very small difference between each of the decile means (about 0.02). Authors note that 48.8 percent (nearly half) of all Providers lowered their scores (i.e., improved)—a coin flip. What would be a more powerful argument is to show the percent of Providers that lowered their scores based on their initial (i.e., 2017) decile ranking. That is, do poorer performing Providers (i.e., higher decile groups based on high (>1.0) MSPB score) show a higher percentage of lowering their score than do Providers in the lower decile groups with lower (<1.0) MSPB scores.
Panel Member #6: NONE
Panel Member #8: Observed to expected cost ratios were identified for all episodes, as well a certain subsets based on site of care post discharge (with or without post-acute care, with or without SNF, with or without downstream readmission). The mean ratio for episodes with a downstream acute readmission was 1.55 with a standard deviation of 0.89 and interquartile range of 1.07 to 1.85.
Spearman correlations between other quality measures were generally positive, albeit less than +0.50.
Panel Member #9: No concerns

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

Submission document: Testing attachment, section 2b5.
Panel Member #1: N/A
Panel Member #2: N/A
Panel Member #3: Not applicable
Panel Member #4: None
Panel Member #5: Face validity information has no meaning in a recertification, as this is deemed insufficient as a measure of validity after the initial certification.
Panel Member #6: NONE
Panel Member #8: Not applicable
Panel Member #9:

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

Submission document: Testing attachment, section 2b6.
Panel Member #1: None.
Panel Member #2: Given that the MSPB Hospital measure primarily uses Medicare claims data, missing data issues are minimal. Other potential issues have been addressed well by the developer.
Panel Member #3: None
Panel Member #4: None
Panel Member #5: Developers state that they expect a "high degree of data completeness". An empirical analysis of >6 million episodes dropped during the analyses used to support the reliability and validity of the measure due to missing data would be more persuasive.
Panel Member #6: NONE
Panel Member #8: Missing data rates were comparable to other Medicare claims based measures and provided. The average O/E ratio was lowest for those with missing DOB and when death date occurred before the trigger date.

16. **Risk Adjustment**

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

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

Panel Member #5: ZIP code level—Area Deprivation Index (ADI) from Census data (2009-2013)

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

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

16d. **Risk adjustment summary:**

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

Panel Member #3: Worth noting, the authors state:

The risk adjustment model now also includes an indicator for whether an episode's index admission was triggered within the 30 day post discharge period of another inpatient stay – to better predict the higher cost of readmission stays (Section 2b3a.3a provides more detail).

Although this is not directly as 'within care' measure, this concept or recent hospitalization basically signals that an admission is a 'readmission' for some other event. Readmissions often cost less than the original admission, but also indicate 'failures' in the initial discharge. In the risk model, these stays are $2,331 more expensive than the grand mean of $9,719. This variable may be necessary to improve model fit, but it also seems to 'forgive' error in the original admission. A potentially better approach would be to count the readmission as part of the total cost of care.

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☐ No

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

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

16d.5. Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☒ No

Panel Member #3: I would say yes, noting my comments above. Also worth noting, I have some lingering concerns about excluding all of the social risk factors. This seems to be a situation where it may be worth paying a high prices for those who qualify by virtue of a disability, for example. However, the analysis suggests these factors are already captured by other things in the model.

Panel Member #5: See previous comments

Panel Member #6: Yes applies to clinical risk factors, NO SES factors were included.

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

Panel Member #1: Standard CMS approach of including DRG and HCCs**.** There are separate models or fully interacted models for each MDC, and tests to assess including interactions.

Risk adjustment r-square varies by MDC with 17 of 26 exceeding 0.25. four are below 0.2,

- Diseases & Disorders Of Blood, Blood Forming Organs, Immunologic Disorders
- Factors Influencing Health Status & Other Contacts With Health Services
- Mental Diseases & Disorders
- Alcohol/Drug Use & Alcohol/Drug Induced Organic Mental Disorders

Model seems to perform well.

SRF does not add to differentiation of risk across patients or hospitals. For hospital scores, most extensive SRF model is correlated with base model excluding all SRF variables at 0.997 (Testing appendix Table 2.b34b.d), and predicted costs when SRF variables included in modeling vary little from base model (Testing appendix Table 2.b34b.c)

Panel Member #2: The measure is modeled through ordinary least square regressions. This measure is defined in rather non-linear ways: *"The numerator for a hospital's MSPB Hospital measure is the average ratio of observed episode cost to expected episode cost across all episodes from a hospital, multiplied by the average observed cost from all hospital episodes nationwide. The numerator is also referred to as the MSPB Hospital Amount. The denominator for a hospital's MSPB Hospital measure is the episode-weighted median MSPB Hospital Amount across all hospitals nationally."*
As indicated above, the measure is non-linear combination of costs. I don't see a rationale that justifies the use OLS as the model of choice for risk-adjustment. Are the underlying random errors normally distributed to justify OLS?
Panel Member #4: Social risk factors are well conceptualized
Panel Member #5: Risk adjustment was generally adequate. Presentation of results in Testing Form (2.b.2) by stratification based on expected higher vs. lower clinical patient costs were generally persuasive that the measure was adequately adjusted for patient-level differences.
Panel Member #6: The risk adjustment model followed the CMS HCC risk adjustment methodology used in Medicare Advantage, including 79 HCC risk factors derived from claims 90 days prior to episode start date. This is **somewhat concerning as 90 days is not typically sufficient time to see all patient's chronic conditions** documented, it typically requires a 12 month look back to see all diagnoses giving time for patient to see their doctor and document all conditions. Other covariates include disability/ESRD status, recent long term care stay, and interaction terms to account for higher cost of some combinations of comorbidities cost more, and evidence of prior admission within past 30 days.
Developers make a strong conceptual argument for including SES, and tested social risk factors by analyzing the impact of the following beneficiary-level and Census-Block Group-level social risk factors: income, education, employment, race, sex, dual status, ADI, and AHRQ Index. These factors were only tested as incremental add ons in step-wise manner to the clinical risk factors already in the model. The model-specific T-tests and partial F-tests (relative to Model 1) indicated that **social risk factors are predictive factors for determining resource use** among beneficiaries for the relevant characteristic and MDC. For example, in models that include the AHRQ SES Index) and models that include the Area Deprivation Index, these indices have p-values less than or equal to 0.05 in 10 of the 26 stratifications. However, the direction of the social risk factor effect was not consistent. **I believe this is due to the imprecise granularity of the data at the Census Block Group level** (only 220K geographic areas that will contain a wide range of high income and low income people for example, which will average out any effects of low income on an outcome like episode cost). They also analyzed the impact of adding social risk variables on overall model performance by looking at the differences in the O/E cost ratio with and without social factors in the risk adjustment model. When including social risk factors in the models, they found minor differences in the O/E ratios. Again, with over 100 variables in the model, this is not a surprising result. They also found results with and without social risk factors were highly correlated, which is also not surprising. They also found provider level effects associated with the social risk factors and did not want to "mask quality".
The model as specified does have good discrimination properties based on clinical risk adjustments applied, though the R-squared values ranged from .11 to .67 across the MDCs, adjusted R-square was similar. The average expected cost differed from the average observed episode cost by 0.06 percent to 1.09 percent in absolute value across deciles. Further, both the predictive and O/E cost ratios were close to one, ranging from 0.99 to 1.01 across risk deciles. These results indicate that the model is accurately predicting spending, regardless of overall risk level.
Panel Member #8: The risk-adjustment model is one commonly used in the assessment of Medicare claims measures, estimated separately for each MDC, determined by the MS-DRG of the index admission and group by principal diagnosis or procedures. An inclusion is a prior inpatient admission risk adjustor, for those admission within the previous 30 days.
Social risk factors are included and comprise dual eligibility, race, sex, and SES from income, education and employment status and zip code.

Panel Member #9: Thorough analysis and sound rationale for utilizing the CMS-HCC model to select risk factors

For cost/resource use measures ONLY:

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

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

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

Panel Member #1: Some concern about excluding patients who died. Ongoing issue of using standardized prices as a measure of resource use.

Panel Member #2:

1. Discharges occurring on Jan 1 or during the early part of measurement period will likely be for admissions that might have occurred the prior year. The allowed Medicare reimbursement rates/amounts may change between those two years – how were that adjusted?

2. The attribution of a readmission that occur in a hospital (say, Hospital B) within the 30-day window following discharge from the index hospital (say, Hospital A). If I understood it correctly, this admission will generate a new triggering event for Hospital B, and the cost associated with this admission will be added to hospital A's cost.

Is that really fair to hospital A? What if hospital B provides lot of additional services that may be considered unnecessary if the patient would have been readmitted to hospital A, given that hospital A already has history of this patient (See S.7.3)?

Panel Member #5: Comparison of "risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital" assumes that the risk adjustment levels the "cost playing field" for potentially widely different hospitals (e.g., urban, rural, private, university-based teaching) and widely different patient populations (e.g., indigent/homeless, high/low dual eligible patient populations, patients with complex surgical needs, patients with complex chronic care needs) can be made equivalent based on the risk adjustment. That is a very tall order. Section S.7.2 provides a detailed description of the Construction Logic, but Step 5 seems to fall short of including these potentially large differences among hospitals in their calculation of expected costs. The results described in Testing Form (2b3.1.1) indicate 109 risk factors including at least some of the potential difference-makers identified previously.

Panel Member #8: No new or usual threats. The assumptions are those common to claims based and registry data applied to participants.

Panel Member #9: No concerns

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: Face validity was assessed by expert panel and review of public comments. Empirical testing consisted of comparing costs of episodes with and without post-acute services expected to increase cost.

Comparison of measure to other cost-specific and efficiency related measures and measures in other HVBP program domains.

Panel Member #2: Both face validity and empirical validity were conducted.

A technical expert panel (TEP) comprising 20 members from diverse backgrounds, including clinicians, healthcare providers, academia, and patient advocacy organizations were assembled, which discussed potential refinement of the MSPB Hospital measure that currently in use from prior rule comments, past

NQF endorsement cycles and other related measure development (e.g., MSPB Clinician). Though no official vote was taken, panelists agreed that maintaining MSPB Hospital measure's holistic "all-cost" approach, allowing readmissions to trigger new MSPB Hospital episodes to increase measure surveillance, and updating the MSPB Hospital measure's MSPB Amount (score numerator) calculation to evenly weight all of a hospital's episodes were appropriate refinements.

Empirical validity was assessed through examining following relationships:

1. Relationship between risk-adjusted episode cost ratios and episodes with and without post-admission events that are known indicators of high cost or intensive care.
2. Relationship between a hospital's average expected episode cost (the average "E" in O/E cost ratios) and average episode rates of several service use categories.
3. Relationship between the MSPB Hospital measure and other cost-specific measures, efficiency-related measures, and measures in other HVBP program domains.

Panel Member #3: Face validity and several different forms of criterion related validity – face validity is helpful; the most compelling pieces of evidence was the O/E ratio for the measure split out by post-episode clinical events. The correlation with other measures was interesting – a smaller number of measures with a tighter logical relationship to the MSPB would make an even stronger case.

Panel Member #4: Face validity testing should be both transparent and systematic. I do not think the process used meets those criteria.

Panel Member #6: The developers used the following methods to test reliability:

1. Face validity was evaluated using a technical expert panel of 20 members with expertise in cost measure development. No official vote was taken, but developers indicate panelists agreed to the hospital measures "all cost" approach, allowing readmissions to trigger new hospital episodes, and updating the numerator calculation to evenly weight all episodes.
2. Empirical Validity Testing: Developers used 3 approaches. First, they examined the relationship between risk-adjusted episode cost ratios and episodes with and without post-admission events known to be indicators of high cost (i.e., observed to expected ratios of episodes with acute care readmissions, with any PAC facility use, and with SNF use. Second, they examined the relationship between a hospital's average expected episode cost and average episode rates of several service use categories. Third, they examined the relationship between the MSPB Hospital measure and other cost-specific measures, efficiency-related measures, and measures in other HVBP program domains, specifically the condition specific Medicare cost measures and with ED wait times.

Panel Member #8: Face validity was derived from TEP, prior rule comments, past NQF endorsement cycles and related measure development.

Empirical validity testing used three approaches. First O/E ratios were calculated for other predictors of high cost such as readmissions, post-acute care facility usage, and SNF care usage. Secondly, correlation with other high cost service usage was performed. Thirdly, correlation with other HVBP measures related to cost or efficiency was performed. As stated previously, higher ratios were obtained for admission with downstream readmissions, post-acute care usage, and SNF usage. Also, Spearman Correlation coefficients were moderately positive correlated with other quality cost measures.

Panel Member #9: TEP and public comments used for face validity. Empirical testing done by assessing costs related to post-admission events related to higher cost of care. In addition, measure was correlated with other MSPB measures in VBP programs.

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

Submission document: Testing attachment, section 2b2.3

Panel Member #1: Expected correlations were observed, with large higher observed to expected when episode had readmission, post-acute care and SNF. (Table 3 in Testing form, p. 19)

Correlation with HVBP clinical outcomes low, but correlation with payment and value of care measures low to moderate, as was the case with correlation to timely and effective care. (Table 4 in Testing form, p. 20)

Panel Member #2: As explained very well in 2b2.3, the correlations/associations were on expected lines indicating empirical validity of the measure.
Panel Member #3: O/E ratios show a clear difference for patients with high cost, post-discharge trajectories. It's harder to know how to interpret the range of correlation statistics between MSPB and other hospital compare measures. In general, they seem to head in the right direction (i.e., positive correlation). However many are small (which, as the developers point out is to be expected).
Panel Member #4: The degree of consensus was moderate to low.
Panel Member #6: The mean, standard deviation, and percentile distribution of observed to expected episode cost ratios for episodes with high-cost post-admission events were higher than their counterparts as expected. For example, episodes with an acute care rehospitalization an average O/E ratio of 1.55 and an interquartile range of 1.07 to 1.85, while episodes without such readmissions had an average O/E ratio of 0.89 and an interquartile range of 0.60 to 1.02. Most service use/setting categories were moderately and positively correlated to the average predicted episode cost, with the correlations across all services categories average +0.487 and procedure use evidencing the strongest correlation +0.721. All three Payment & Value of Care measures, capturing 30-day Medicare payments for acute myocardial infarction, heart failure, and pneumonia conditions, were positively but weakly correlated with the hospital average predicted episode cost.
Panel Member #8: Besides face validity, the validity, as stated by the developers, obviously correlates with other costly subsequent or high dollar procedures. This may seem like a syllogism, but does predict the episode cost of the current admission fairly reliably and correlates with downstream high cost episodes.
Panel Member #9: Demonstrated low to moderate correlation between costs and other unplanned events

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

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☐ **No**

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

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

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

☒ **Yes**

☐ **No**

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

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

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

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

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

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

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

Panel Member #1: Face validity high. Exclusions reasonable, although excluding patients who died needs more justification beyond reducing variability. Risk adjustment model performs reasonably.
Panel Member #2: I am giving a moderate rating based on my notes in # 18.

Panel Member #3: In general, I think it is hard to validate claims based cost measures. I thought the O/E ratio analysis was convincing, but limited in terms of all the different dimensions of validity that could be taken into consideration. Its broad use is another form of validation – would love to know if the measure has any predictive validity.
Panel Member #4: There is something sort of de facto valid about a cost and resource use measure that accounts for non-behavior related determinants of cost (e.g. standardized prices)
Panel Member #5: Developer demonstrated an effort to risk adjust measure to create valid measure score.
Panel Member #6: Validity results were based on several approaches and results were in hypothesized direction, but most were not strong. Face validity results were not provided.
Panel Member #8: Again, the value is to predict the cost of the current episode, which would be expected to be correlated with comorbidities, procedures and services rendered during that episode and subsequent care delivered post discharge. In a way, this measure not only seems to confirm the obvious, but provide some ranges of the ratios for various groups, which could hopefully drive performance improvement to reduce the cost of the current admission.
Panel Member #9: Demonstrated correlation, would have rated higher if correlation was stronger but submitters did expect lower results and provided rationale

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 #8: Does it provide useful and actionable information?

Brief Measure Information

NQF #: 2158

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

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

De.3. Brief Description of Measure: The MSPB Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital. Specifically, the MSPB Hospital measure assesses the cost to Medicare for Part A and Part B services performed by hospitals and other healthcare providers during an MSPB Hospital episode, which is comprised of the periods 3-days prior to, during, and 30-days following a patient's hospital stay. The MSPB Hospital measure is not condition specific and uses standardized prices when measuring costs. Beneficiary populations eligible for the MSPB Hospital calculation include Medicare beneficiaries enrolled in Medicare Parts A and B who were discharged between January 1 and December 1 in a calendar year from short-term acute hospitals paid under the Inpatient Prospective Payment System (IPPS).

IM.1.1. Developer Rationale: The MSPB Hospital measure is included in the Efficiency and Cost Reduction domain of the Hospital VBP program. With measures in other domains of clinical outcomes, safety, and person and community engagement, the HVBP program provides financial incentives to hospitals to further the value of care they provide.

The MSPB Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital. This scoring allows hospitals to improve their score by spending less than the episode-weighted risk-adjusted median cost during a given performance period through improved care coordination and provision of efficient care. For instance, hospitals can decrease (i.e., improve) their risk-adjusted episode costs through actions such as:

1) improving coordination with post-acute providers to reduce the likelihood post-discharge of adverse events,
2) identifying unnecessary or low-value post-acute services and reducing or eliminating these services, or
3) shifting post-acute care from more expensive services (e.g., skilled nursing facilities) to less expensive services (e.g., home health) in cases that would not affect patient outcomes. Care coordination helps ensure a patient's needs and preferences for care are understood, and that those needs and references are shared between providers, patients, and families as a patient moves from one healthcare setting to another. People with chronic conditions, such as diabetes and hypertension, often receive care in multiple settings from numerous providers. As a result, care coordination among different providers is required to avoid waste, over-, under-, or misuse of prescribed medications and conflicting plans of care.

De.1. Measure Type: Cost/Resource Use

S.5. Data Source: Assessment Data

Claims

Enrollment Data

Other

S.3. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Dec 09, 2013 **Most Recent Endorsement Date:** Jul 13, 2017

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? N/A

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)

The MSPB Hospital measure is included in the Efficiency and Cost Reduction domain of the Hospital VBP program. With measures in other domains of clinical outcomes, safety, and person and community engagement, the HVBP program provides financial incentives to hospitals to further the value of care they provide.

The MSPB Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital. This scoring allows hospitals to improve their score by spending less than the episode-weighted risk-adjusted median cost during a given performance period through improved care coordination and provision of efficient care. For instance, hospitals can decrease (i.e., improve) their risk-adjusted episode costs through actions such as:

1) improving coordination with post-acute providers to reduce the likelihood post-discharge of adverse events,
2) identifying unnecessary or low-value post-acute services and reducing or eliminating these services, or
3) shifting post-acute care from more expensive services (e.g., skilled nursing facilities) to less expensive services (e.g., home health) in cases that would not affect patient outcomes. Care coordination helps ensure a patient's needs and preferences for care are understood, and that those needs and references are shared between providers, patients, and families as a patient moves from one healthcare setting to another. People with chronic conditions, such as diabetes and hypertension, often receive care in multiple settings from numerous providers. As a result, care coordination among different providers is required to avoid waste, over-, under-, or misuse of prescribed medications and conflicting plans of care.

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.**

Analysis of all IPPS eligible hospitals with at least 25 episodes for the 2018 performance period

shows a large range of provider scores on the MSPB Hospital measure. The MSPB Hospital measure score has the following distributional characteristics:

- Mean: 0.99, standard deviation: 0.08
- Median: 0.99
- Min: 0.49, max: 1.68
- Interquartile range spans from 0.94 to 1.03

The score decile distribution for the 2018 performance period is:

- 10th: 0.90
- 20th: 0.93
- 30th: 0.95
- 40th: 0.97
- 50th: 0.99
- 60th: 1.01
- 70th: 1.02
- 80th: 1.05
- 90th: 1.08

Analysis of MSPB Hospital measure score changes between 2017 and 2018 showed that hospital scores do vary over time, as 48.8 percent of providers evidenced improved (lower) scores. The distribution in score change between these two years, with negative values indicating improvement, is

- Min: -166.24%
- 5th: -17.54%
- 10th: -4.15%
- 25th: -1.76%
- 50th: 0.10%
- 75th: 2.01%
- 90th: 4.41%
- 95th: 18.92%
- Max: 35.68%

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.

The response to IM2.2 includes measure scores calculated for all IPPS-eligible hospitals with at least 25 episodes during the performance period of January 1, 2018 to December 1, 2018.

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.**

We analyzed disparities data through analysis of beneficiary and community/regional characteristics based on or directly from the American Community Survey (ACS) and CMS' Enrollment Database (EDB). All ACS variables firstly defined at the Census Block Group level and then ZIP code when census block group is missing. The specific social risk factors (SRFs) analyzed include the following variables.

- Income (ACS): Low Income: median income < 33rd percentile nationally; Medium Income: median income in the interval spanning the 33rd percentile to the 66th percentile nationally; High Income: median income > 66th percentile
- Education (ACS): Education < High School: when % with < high school education is the highest for a given Census Block Group; Education = High School: when % with only high school is the highest; Education > High School: when % with > high school is the highest
- Employment (ACS): Unemployment Rate > 10%; Unemployment Rate <= 10%
- Race (EDB): Asian, Black, Hispanic, North American Native, White, and Other
- Sex (EDB): Female, male
- Dual status (CME): Full dual, partial dual, non-dual status
- Area Deprivation Index (ADI)[1]
- Agency of Healthcare Research and Quality (AHRQ) SES Index: AHRQ index scores are calculated using the AHRQ scoring algorithm and is a continuous dependent variable as a replacement of all SES variables. The index includes percentage of households containing one or more person per room, median value of owner-occupied dwelling, percentage of persons below the federally defined poverty line, median household income, percentage of persons aged = 25 years with at least 4 years of college, percentage of persons aged = 25 years with less than a 12th grade education, and percentage of persons aged 16 or older in the labor force who are unemployed.[2]

Out of 4,023,571 beneficiaries and 5,984,315 beneficiary episodes across all major diagnostic categories [3], the percentage of female beneficiaries range from 27.0 percent to 63.5 percent across the 23 of the 26 MDCs in this measure that reasonably occur for both sexes (MDC 13 and MDC 14 are nearly 100 percent female as they are related to pregnancy, childbirth, and the female reproductive system, while MDC 12 is 0 percent female as it is related to the male reproductive system). For 23 out of 26 MDCs, most beneficiaries (55.7% - 84.4%) have non-dual status. The MDCs with a minority of non-dual status beneficiaries includes MDC 14 – Pregnancy, Childbirth, and the Puerperium (12.1%), MDC 25 – Human Immunodeficiency Virus Infections (30.1%), and MDC 19 – Mental Diseases and Disorders (44.0%). Income level is categorized into high, medium, and low from the continuous average income variable in ACS; therefore, each category has 33.3 percent of episodes. Approximately 2.0 to 8.1 percent of beneficiaries across all MDCs are classified as having below a high school education level, while 16.8 to 37.1 percent of beneficiaries have high unemployment designation (>10% for the Census Block Group). The AHRQ Index ranged from 28.82 to 78.4 across beneficiary episodes and approximately 14.36 of beneficiary episodes were ranked in the top quintile of the ADI's national ranking.

We also analyzed the effect and impact of several social risk factors in the MSPB Hospital measure's risk adjustment model and sought to determine the extent to which these effects may be attributable to hospitals versus the patients they serve. As in our previous studies, we found inconsistency in the beneficiary-level estimates of the social risk factors and minimal impact to MSPB Hospital scores. Moreover, we found statistically significant hospital-level effects when decomposing the effects of select social risk factors between hospitals and beneficiaries.

[1] University of Wisconsin School of Medicine Public Health. 2015 Area Deprivation Index v2.0. Downloaded from https://www.neighborhoodatlas.medicine.wisc.edu/ February 24, 2020.

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

[3] Note that SRF testing occurred over a smaller set of beneficiary episodes than most other testing as approximately 1.7 percent of beneficiary episodes with missing income/employment ACS data were excluded from SRF studies.

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

N/A

IM.2. Measure Intent

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

The MSPB Hospital measure aims to incentivize hospitals to coordinate care and reduce unnecessary utilization during the period immediately prior to, during, and in the 30 days after a hospital discharge. Because a hospital's MSPB Hospital measure score is based on all Medicare Part A and Part B claims data for episodes during the period of performance and is not condition-specific, the MSPB Hospital measure evaluates hospitals' efficiency across all conditions and admissions. The all-cause nature of the MSPB Hospital measure makes the measure relevant to a large number of hospitals, maximizing its impact. The effect of patient health status and demographics on episode spending is accounted for by the MSPB Hospital's risk adjustment methodology. One can measure whether hospitals provide efficient care by examining the MSPB Hospital measure alone as well as in concert with a variety of quality of care measures already reported on CMS' Hospital Compare webpage and developed as part of CMS's Hospital VBP Programs.

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):*

Care Coordination

Safety : Overuse

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

Inpatient/Hospital

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

<WebPageURLExists nodeType="1">https://www.qualitynet.org/files/5f1b3bd12bd4670021abc1b4?filename=MSPB_Hospital_MIF_2020.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.)*

Medicare Part A and Part B claims data: Part A and B claims data are used to build MSPB Hospital episodes, calculate episode costs, and construct risk adjustors. CMS Office of Information Systems (OIS) maintains a detailed Medicare Claims Processing Manual available at the following URL: https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Internet-Only-Manuals-IOMs-Items/CMS018912.

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

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

We used additional data sources for measure testing purposes:

- American Community Survey (ACS): This is used for evaluating social risk factors. https://www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html.
- Common Medicare Environment (CME) database: This is used for evaluating social risk factors. https://www.ccwdata.org/documents/10280/19002256/medicare-enrollment-impact-of-conversion-from-edb-to-cme.pdf.
- Area Deprivation Index (ADI): University of Wisconsin School of Medicine Public Health. 2015 Area Deprivation Index v2.0. Downloaded from https://www.neighborhoodatlas.medicine.wisc.edu February 24, 2020.

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)*

2020-07-27-nqf-testing-appendix-mspb-hospital-v4.xlsx

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

Data Dictionary:

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

Please supply the username and password:

Attachment: S_6_Data_Dictionary-637425096966930043.xlsx

Code Table:

URL:

Please supply the username and password:

Attachment: S_6_Code_Table.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 Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital.

The MSPB Hospital measure methodology first identifies hospital discharges occurring between January 1 and December 1 of a calendar year and that occur at acute care hospitals paid under the Medicare's IPPS. A set of exclusion criteria, detailed in Sections S.7.2 and S.9.1, are applied to these discharges to promote measure population comparability.

The measure methodology then defines MSPB Hospital episode timeframes, which span from the 3-days prior to a hospitalization, a hospitalization period, and 30-days following hospitalization discharge.

Third, all Medicare Part A and Part B standardized costs for services initiated during an MSPB Hospital episode are then summed to provide the total observed episode cost and risk-adjusted to provide the total expected episode cost.

Finally, MSPB Hospital measure is calculated for each acute care IPPS hospital. The numerator for a hospital's MSPB Hospital measure is the average ratio of observed episode cost to expected episode cost across all episodes from a hospital, multiplied by the average observed cost from all hospital episodes nationwide. The numerator is also referred to as the MSPB Hospital Amount. The denominator for a hospital's MSPB Hospital measure is the episode-weighted median MSPB Hospital Amount across all hospitals nationally.

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

STEP 1: Define and Trigger Episodes

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

STEP 2: Standardize Claim Payments

Medicare Part A and B costs occurring during episodes are standardized to promote cost comparability while preserving differences that result from healthcare delivery choices. This standardization process, also referred to as payment standardization, adjusts the allowed charge for services by removing geographic differences (e.g., due to labor costs) and adjustments from special Medicare programs (e.g., graduate medical education and disproportionate share payments).

Payment standardization is applied to several measures, including the MSPB Hospital measure, and is detailed at https://www.resdac.org/articles/cms-price-payment-standardization-overview

STEP 3: Apply Exclusion Criteria

Exclusions that are based on beneficiary or hospitalization characteristics are applied to promote episode comparability and completeness. Episodes are excluded from the MSPB Hospital measure if they meet any of the following conditions:

- The beneficiary has a primary payer other than Medicare during the episode window or in the 90-day lookback period
- Beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 90-day lookback period and episode window
- The beneficiary's death occurred during the episode.
- The index admission for the episode did not occur in a subsection (d) hospital paid under the Inpatient Prospective Payment System or occurred in a Maryland hospital.
- The index admission for the episode is involved in an acute-to-acute hospital transfer (i.e., the admission ends in a hospital transfer or begins because of a hospital transfer).
- The index admission inpatient claim indicates a $0 actual payment or a $0 standardized payment.

STEP 4: Calculate Observed Episode Cost

Observed episode cost is the sum of all the standardized Medicare claims payments (allowed amounts) for services initiated during the MSPB Hospital episode, between 3-days prior to the hospital admission until 30-days after discharge.

The costs for Medicare Part A and B services that are initiated during an episode and extend in duration beyond the episode are not prorated. Thus, for example, if a patient begins Inpatient Rehabilitation Facility (IRF) care within 30-days of discharge from an index admission, then the episode will contain the full Medicare cost of that IRF claim.

STEP 5: Calculate Expected Episode Cost

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

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

- Define HCC and patient characteristic-related risk adjustors using Medicare Parts A and B claims in the 90-day lookback period from the episode start date.
- Define other risk adjustors that rely upon Medicare beneficiary enrollment and assessment data as follows:

  - Identify beneficiaries who are originally "Disabled without end-stage renal disease (ESRD)" or "Disabled with ESRD" using the original reason for joining Medicare field in the Medicare beneficiary enrollment database.
  - Identify beneficiaries with ESRD if their enrollment indicates ESRD coverage, ESRD dialysis, or kidney transplant in the Medicare beneficiary enrollment database in the 90-day lookback period.

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

- Categorize beneficiaries into age ranges using their date of birth information in the Medicare beneficiary enrollment database.
- Calculate an ordinary least squares (OLS) regression model to estimate the relationship between all the risk adjustment variables and the dependent variable, the standardized observed episode cost, to obtain the expected episode cost. A separate OLS regression is run for each episode MDC group nationally.
- Winsorize the expected episode cost by assigning the value of expected episode cost at the 0.5th percentile of the distribution for episodes within the same MDC to all episodes with expected episode costs below the 0.5th percentile.
- Renormalize values by multiplying each episode´s winsorized expected cost by the ratio of the MDC group´s average observed cost and the MDC group´s average winsorized expected cost.
- Exclude episodes with outlier residuals to obtain finalized expected episode cost. This step is performed across all episodes regardless of the MDC group.

- Calculate each episode´s residual as the difference between the observed cost and the re-normalized, winsorized expected cost computed above.
- Exclude episodes with residuals below the 1st percentile or above the 99th percentile of the residual distribution.
- Renormalize all remaining episodes by multiplying their cost by the ratio of the average observed episode cost and the average winsorized expected cost when excluding outliers.

STEP 6: Calculate Measure Scores

The MSPB Hospital measure is calculated for each hospital as the average ratio of observed episode cost to expected episode cost across all episodes from that hospital, multiplied by the average observed cost from all hospital episodes nationwide. The numerator is also referred to as the MSPB Hospital Amount. The denominator for a hospital's MSPB Hospital measure is the episode-weighted median MSPB Hospital Amount across all hospitals nationally.

The MSPB Hospital measure methodology presented in this Intent to Submit form and accompanying Testing Attachment differs from the methodology previously endorsed by NQF in 2016 and that is in current use in CMS programs in three ways. First, the MSPB Hospital Amount, as calculated in Step 6, now imposes equal weight to all risk-adjusted hospital episodes by using the average ratio of observed to expected episode costs instead of the ratio of average observed episode costs to average expected episode costs. Second, the refined measure presented in this form expands the coverage of episodes included in the MSPB Hospital measure by allowing acute care re-hospitalizations that occur within 30-days of a hospital discharge to trigger MSPB episodes (Step 1). While the cost of such readmission events were captured in the original methodology, [1]they were not permitted to initiate new MSPB Hospital episodes. Third, the refined methodology adds into the risk adjustment process (Step 5) a control variable that accounts for these newly triggered admissions occurring within 30 days of another index hospitalization discharge date, ensuring that a hospital's risk-adjusted episode cost on these newly triggered episodes is accurately estimated.

[1] Specifically, in the original MSPB Hospital methodology, the cost of such a readmission event would be captured in the preceding index admission's 30 day post-discharge period.

S.7.2a. CONSTRUCTION LOGIC ATTACHMENT or URL: If needed, attach supplemental documentation (Save file as: S_7_2_Construction_Logic). All fields of the submission form that are supplemented within the

attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.

URL: See URL provided in Section S.1

Please supply the username and password:

Attachment:

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

The MSPB Hospital measure includes Medicare Part A and Part B services that are furnished to a beneficiary during the episode. The MSPB Hospital measure avoids redundancy of clinical events by counting each service once within an episode.

The MSPB Hospital measure allows episode overlap in cases of acute care hospital readmissions. Example: Consider a patient who is discharged from Hospital A and is admitted to Hospital B within 30 days of discharge from Hospital A. The first hospitalization would trigger an MSPB Hospital episode that is attributed to Hospital A and the second hospitalization would trigger an MSPB Hospital episode that is attributed to Hospital B. As that second hospitalization occurred within 30-days of discharge from the first hospitalization, the cost of the second hospitalization would be included as part of Hospital A's MSPB episode cost. The cost of the second hospitalization is also included in Hospital B's episode. As such, the second hospitalization is counted only once in each episode and allows the MSPB Hospital measure to ensure continuous accountability among providers and throughout a beneficiary's trajectory of care. As noted in Section S.7.1 of this document, the MSPB Hospital measure methodology, as previously endorsed by NQF in 2016 and as currently used by CMS, did not allow the second hospitalization in this example to trigger a new episode.

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

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

An episode includes all services from the 3 days prior to a hospital admission to promote MSPB Hospital episode consistency in potentially complementary services, regardless of the diagnosis code or type of pre-admission services that may occur.

Specifically, diagnostic services and non-diagnostic services related to the reason for admission are captured in the inpatient diagnosis-related group (DRG) payment for the hospitalization when they are performed by the hospital during the 3 days prior to admission. However, diagnostic services or non-diagnostic services related to the reason for admission that are performed by a provider other than the hospital are not captured in the inpatient DRG payment and are paid separately under Medicare. Furthermore, non-diagnostic services that appear to be unrelated to the reason for admission are also not captured in the inpatient DRG payment and are paid separately under Medicare. The MSPB Hospital episode includes all services from 3 days prior to ensure that all costs are included in the measure. For additional discussion, please refer to S.8.4., which details the rationale for the construction of the MSPB Hospital episode.

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

Clinical hierarchies are embedded in the risk adjustment model, described in Section S.7.2 and in more detail in Sections S.8.4 and S.8.5. The MSPB Hospital measure uses variables from CMS' Hierarchical Condition Category

(HCC) model. This approach is adopted to ensure sufficient capture of the patient's comorbid disposition prior to the index hospital admission and allow more comprehensive risk adjustment of comorbid factors. The model suppresses HCCs for less severe manifestations of a conditions when evidence for the more severe condition is found to prevent collinearity in regression estimation.

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

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

CMS has in place several auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and to recoup any overpayments. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in this measure, including diagnosis and procedure codes and other elements that are consequential to payment. Specifically, CMS works with Zone Program Integrity Contractors (ZPICs), and formerly Program Safeguard Contractors (PSCs), to ensure program integrity; the agency also uses Recovery Audit Contractors (RACs) to identify and correct for underpayments and overpayments.

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

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

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

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

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

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

Inpatient services: Inpatient facility services

Inpatient services: Evaluation and management

Inpatient services: Procedures and surgeries

Inpatient services: Imaging and diagnostic

Inpatient services: Lab services

Inpatient services: Admissions/discharges

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)

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 Hospital measure assesses the standardized allowed amounts of services during an MSPB episode, which includes all Medicare Parts A and B claims that occur 3 days prior to the index admission through 30 days after the hospital discharge. This identification approach allows the MSPB Hospital measure to capture the breadth of service categories that can be attributed to the hospital where the beneficiary's episode of care was initiated.

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

URL: See URL provided in Section S.1

Please supply the username and password:

Attachment:

Clinical Logic

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

Objective: The MSPB Hospital measure aims to improve care coordination and care quality in the period between 3 days prior to an acute inpatient hospital admission through the period 30 days after discharge.

Clinical Topic Area: Inpatient Admissions, all conditions

Accounting for Comorbidities: Application of a variant of the CMS-HCC risk adjustment model. The model includes a full set of interaction terms between comorbidities and MDC of the index admission, as well as a select number of interaction terms between comorbidities.

Measure of Episode Severity: Risk adjustment model includes indicators for the MS-DRG of the index admission.

Concurrency of Clinical Events. The MSPB Hospital episode spans the period 3 days prior to the index hospital admission through 30-days post-discharge. All Medicare Part A and B claim-based events initiated during this period are included in the MSPB Hospital episode.

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.)*

Objective: The MSPB Hospital measure aims to improve care coordination in the period between 3 days prior to an acute inpatient hospital admission through the period 30 days after discharge. The MSPB Hospital

measure recognizes lower costs associated with a reduction in unnecessary services, preventable complications, readmissions, and shifting post-acute care from more expensive to less expensive services when appropriate.

Grouping methodology: The MSPB Hospital measure evaluates resource use through the unit of MSPB Hospital episodes. The MSPB Hospital episodes are constructed by including all Medicare Part A and Part B claims with a start date falling between 3 days prior to an acute inpatient hospital admission through the period 30 days after discharge. Episodes that may provide an incomplete or non-comparable view of episodes spending, such as when a beneficiary enrolled in a Medicare Advantage plan, are excluded from measure calculation. A full set of exclusion criteria are provided in Section S.7.2.

Cost Calculation: The MSPB Hospital amount includes the cost of services performed by hospitals and other healthcare providers during an MSPB Hospital episode, which is comprised of the period 3 days prior to an inpatient PPS hospital admission (index admission) through 30-days post-hospital discharge. All costs are payment standardized to control for geographic variation in Medicare reimbursement rates. To account for the clinical severity of patients, standardized costs are risk adjusted at the Major Diagnostic Category (MDC) level, using a combination of clinical indicators of CMS' Hierarchical Condition Category Version 22 (CMS-HCC V22) risk adjustment model (patient-level), an indicator of the severity of the index hospitalization (hospital stay, MS-DRG), an indicator of whether an index hospitalization is initiated within 30 days of another inpatient stay, indicators that rely on Medicare beneficiary enrollment and assessment data (patient level, e.g., ESRD coverage), and combinations thereof. The risk adjustment models are run within each MDC and with these indicators to support comparability across episodes. Further, the risk adjustment indicators are assessed over the 90 days preceding the episode to ensure that clinical events occurring near the episode window are captured and to minimize the loss of data for patients with a limited history of Medicare claims and administrative data. The indicators used for risk adjustment and the methodology are detailed in the Measure Information Form linked in Section S.1.

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)*

Grouping Methodology:

The MSPB Hospital measure methodology defines an MSPB Hospital episode as all claims with start dates falling between 3 days prior to an IPPS hospital admission (index admission) through 30-days post-hospital discharge and does not separate concurrent events. It includes services initiated in the period 3-days prior to hospital admission, during the hospitalization, and 30 days after hospital discharge to emphasize the importance of care transitions and care coordination in improving patient care and reducing unnecessary readmissions.

This episode grouping approach is consistent the MSPB Hospital measure's original intent and provides continued value as newer cost measures focus on condition- and procedure-specific episodes of care. Indeed, the MSPB Hospital measure's episode definition is consistent with MedPAC's response to the FY 2012 IPPS proposed rule in which they recommended that "both CMS and MedPAC should focus on creating parallel incentives for hospitals and post-acute care providers to work to reduce readmissions. The end goal is to align incentives across the sectors to encourage cooperation among providers to improve the quality of the episode of care, reduce the cost of the episode of care, and reduce the number of unnecessary inpatient episodes." [1] More recently, in 2016, MedPAC noted their belief that hospitals be "rewarded or penalized based on a broad all-condition 30-day cost measure", that "cost measures used should be as broadly based as possible" to "ensure reliability and provide a broad incentive to reduce costs across all types of services", and their support for the use of the MSBP Hospital measure in CMS programs [2]. This episode grouping approach is also consistent with NQF's theoretical definition of an episode of care in that it is "…a series of temporally

contiguous healthcare services related to the treatment of a given spell of illness or provided in response to a specific request by the patient or other relevant entity."[3]

Cost Calculation:

The inpatient setting is an area of high spending where increased cost effectiveness can be impactful in keeping Medicare spending affordable: in 2016, Medicare FFS paid $183 billion for approximately 10 million Medicare inpatient admissions and 200 million outpatient services, which reflects a 2.3 percent increase in hospital spending per FFS beneficiary between 2015 and 2016 [4]. Of the $190 billion that the Medicare FFS program and FFS beneficiaries paid to 4,700 short-term acute care hospitals in 2018 $121 billion was for inpatient stays – an increase of 1.1 percent from 2017 [5]. Given that the inpatient hospital setting is such an important contributor to overall Medicare spending, it is necessary to measure costs related to hospitalizations.

The MSPB Hospital measure offers opportunity for improvement where providers can exercise influence on costs during the hospitalization or contiguous after care. Through its episode grouping and cost capture, providers can assess the cost of care for patients, identify particularly costly episode characteristics; and, with quality measures, determine the value of care provided to patients. To promote these activities, the clinical logic for the model used to risk adjust episode cost affords equitable patient episode and measure comparisons by controlling for patient clinical characteristics prior to episode start. Patient comorbidities are associated with higher resource use in the inpatient setting, such as through additional hospitalization charges, longer stays, and higher readmission rates. These include comorbidities for chronic conditions; for example, diabetes, hypertension, and heart failure have been found to be associated with higher levels of resource use [6,7]. Also, psychiatric comorbidities (e.g., depression, anxiety, dementia, substance use, bipolar disorders) have been associated with higher readmission rates for common inpatient treatment.[8,9] Medicare beneficiaries with multiple comorbidities account for a disproportionate amount of expenditure, including through additional resource use and length of stays [10,11]. As such, it is important to account for patient comorbidities and disease interactions in a resource use measure.

[1] FY2012 IPPS Final Rule https://www.govinfo.gov/content/pkg/FR-2011-08-18/pdf/2011-19719.pdf

[2] MedPAC Letter to Acting Administrator RE: File Code CMS-1655-P

http://www.medpac.gov/docs/default-source/comment-letters/medpac-comment-on-cms-s-proposed-rule-on-hospital-inpatient-prospective-payment-systems-for-acute-ca.pdf?sfvrsn=0

[3] 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

[4] MedPAC. (2018) Report to the Congress: Medicare Payment Policy."

[5] MedPAC. (2020) Report to the Congress: Medicare Payment Policy."

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

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

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

[8] Sayers, SL., Hanrahan N, Kutney A, Clarke S, Reis BF, and Riegel B. "Psychiatric Comorbidity and Greater Hospitalization Risk, Longer Length of Stay, and Higher Hospitalization Costs in Older Adults with Heart

Failure." Journal of the American Geriatrics Society 55, no. 10 (October 2007): 1585–91. https://doi.org/10.1111/j.1532-5415.2007.01368.x

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

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

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

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

URL: See URL provided in Section S.1

Please supply the username and password:

Attachment:

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

Trigger Event: admission to acute care hospital ("index admission")

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

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

The triggering and ending mechanism allow consistent capture of services initiated during the period directly surrounding an inpatient stay. The static timing of the episode start and end dates and use of all Medicare Part A and B claims minimize the complexity of this measure, making the easily implementable and readily actionable.

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

Services captured 30 days after a hospital discharge emphasize the importance of care transitions and care coordination. The length of this period is long enough to capture costs related to the hospital stay, without being so long as to reduce the attributed providers' influence, aligns with other measures, and corresponds to identified care coordination and cost surveillance needs, as noted in Section S.8.3.

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

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

Comorbidity/Major Complication and Comorbidity. The risk adjustment model also includes an indicator for whether the index admission was triggered within the 30 day post discharge period of another inpatient stay.

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

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

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

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

Adjustments for Comparability

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

:

The MSPB Hospital measure calculation is comprised of Medicare beneficiary episodes of care for beneficiaries and hospitals that do not meet population exclusion criteria. The population exclusion criteria promote comparability across the population captured by this measure. MSPB Hospital measure's risk adjustment, which includes Winsorization for extreme values and outlier exclusion, further promotes measure comparability at its most granular level, the episode level.

Population Exclusions for Comparability.

As discussed in Section S.7.2, Step 3, the MSPB Hospital measure excludes episodes based on select hospitalization or beneficiary characteristics to foster comparability in service use and population captured by the measure. Specifically, the measure excludes episodes that meet any of the following criteria:

- The beneficiary has a primary payer other than Medicare during the episode window or in the 90-day lookback period
- The beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 90-day lookback period and episode window
- The beneficiary's death occurred during the episode.
- The index admission for the episode did not occur in neither a subsection (d) hospital paid under the Inpatient Prospective Payment System (IPPS) or occurred in a Maryland hospital.

- The index admission for the episode is involved in an acute-to-acute hospital transfer (i.e., the admission ends in a hospital transfer or begins because of a hospital transfer).
- The index admission inpatient claim indicates a $0 actual payment or a $0 standardized payment.

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

Statistical Adjustments for Comparability.

The MSPB Hospital measure also applies risk adjustment and statistical exclusions and renormalization to further ensure comparability. These adjustments are fully described in Step 5 of the construction methodology (Section S.7.2). The risk adjustment approach accounts for patient level variation prior to the index hospitalization and the severity of the index hospitalization through regression models. The statistical exclusions and renormalizations that follow cost predictions from these models ensure that cost distributions resulting from outlier exclusions remain true to population averages.

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

S.9.2. Risk Adjustment Type (Select type)

Stratification by risk category/subgroup

If other:

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

The MSPB Hospital measure is stratified by Major Diagnostic Category (MDC), which are mutually exclusive groups of MS-DRGs that correspond to an organ system (e.g., diseases and disorders of the digestive system) or cause (e.g., burns). There are 25 MDCs (numbered 01-25), and a Pre-MDC group for extremely resource intensive MS-DRGs. MS-DRGs within the numbered MDCs are largely determined by principal diagnosis, while MS-DRGs within the Pre-MDC group are determined by Operating Room procedures (e.g., organ transplant).

The MSPB Hospital measure's MDC stratification and risk adjustment model, which controls for episode MS-DRG, allows for equitable patient episode comparisons that preserve clinically meaningful distinctions in the beneficiary population within each MDC.

The risk adjustment variables included in the model are listed in document hyperlinked in Section S.1.

S.9.4 Costing method

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

Standardized pricing

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

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

Ratio

Attachment

If other:

Attachment: S10_sample_score_report.xlsx

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 Hospital measure that is less than 1 indicates that a hospital's MSPB Hospital Amount (i.e. risk-adjusted spending) is less than the national episode-weighted median MSPB Hospital Amount across all hospitals during a given performance period. An MSPB Hospital measure that is greater than 1 indicates that a hospital's MSPB Hospital Amount (i.e. risk-adjusted spending) is greater than the national episode-weighted median MSPB Hospital Amount across all hospitals during a given performance period.

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

As described Step 6 in Section S.7.2, the MSPB Hospital measure is calculated for each hospital as the average ratio of observed episode cost to expected episode cost across all episodes from that hospital, multiplied by the average observed cost from all hospital episodes nationwide. The numerator is also referred to as the MSPB Hospital Amount. The denominator for a hospital's MSPB Hospital measure is the episode-weighted median MSPB Hospital Amount across all hospitals nationally.

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).

The MSPB Hospital measure version under consideration has not been reported under the Hospital Value Based Purchasing Program. The revised MSPB Hospital measure's use in CMS programs, like the Hospital Value-Based Purchasing (VBP) program, is expected after legislative public reporting requirements for the Hospital Inpatient Quality Reporting and HVBP program are met. The version under consideration differs from the previously NQF-endorsed MSPB Hospital measure version that is in current use by CMS programs ("current version") in that the version under consideration allows acute care hospital readmissions to trigger a new MSPB episode and changes the calculation of the MSPB Amount (the measure score numerator) from a calculation based on the ratio of average observed episode cost to average expected episode cost to an average ratio of observed to expected episode cost (see the end of Section S.7.1 for more detail).

The distribution of all MSPB Hospital measure scores for in 2018 between both measure versions are provided below (current version versus revised version).

For all hospitals with an MSPB Hospital measure, the distribution is:

- Maximum : 2.03 vs. 2.00
- 90th percentile: 1.08 vs. 1.09
- 75th percentile: 1.03 vs. 1.03
- 50th percentile: 0.99 vs. 0.99
- 25th percentile: 0.94 vs. 0.94
- 10th percentile: 0.89 vs. 0.89

- Minimum : 0.31 vs. 0.32

And, for all hospitals with at least 25 episodes, the distribution is:

- Maximum : 1.53 vs. 1.68
- 90th percentile: 1.08 vs. 1.08
- 75th percentile: 1.03 vs. 1.03
- 50th percentile: 0.99 vs. 0.99
- 25th percentile: 0.94 vs. 0.94
- 10th percentile: 0.89 vs. 0.90
- Minimum : 0.48 vs. 0.49

A distribution of hospitals' MSPB measure values is provided to hospitals as part of their hospital-specific reports (HSRs). As noted in Section S.7.2., the denominator of the MSPB Hospital measure is weighted by the number of episodes; as a result, the (unweighted) median MSPB Hospital measure score is not necessarily always equal to one.

The MSPB Hospital measure is also reported to hospitals with information about the national average measure and the state average measure for the specific state that the hospital is a part of. Hospitals can also see the national and state average observed and expected spending per MDC and the national and state percent of spending for each claim type within the episode window. With this information, hospitals can identify the areas where the observed and expected spending are most concentrated and is most different from the national and state average.

Because CMS uses the full population of Medicare Parts A and B claims data to calculate the MSPB Hospital measure and due to the large sample sizes, confidence intervals are of limited value. The calculated MSPB Hospital measure represents the true measure for the period of interest. A confidence interval is still of value in assessing the "statistical noise" in a hospital's measure score,

but the reliability metrics presented in this submission also formally assess the extent of "statistical noise" and the ability to distinguish between providers' performance.

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.

An MSPB Hospital episode is attributed to the hospital whose inpatient admission triggered the episode (Section S.8.4).

Hospitalizations eligible to start an MSPB Hospital episode must end in a discharge 30 days prior to the end of the period of performance to permit the collection of claim information during the post-discharge period. Further, as noted in S.9.1., acute-to-acute hospitalization transfers are not eligible to trigger an episode due to the uncertainty surrounding proper attribution of such episodes.

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.

All short-term acute inpatient prospective payment system (IPPS) hospitals. Short-term acute IPPS hospitals are hospitals in the 50 States and D.C. other than: psychiatric hospitals, rehabilitation hospitals, and long-term care hospitals. The measure also excludes inpatient facilities whose patients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer. [1]

[1] The MSPB Hospital uses the CMS definition of a cancer hospital: http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/PPS_Exc_Cancer_Hospasp.html

S.13.4. Sample size

Detail the sample size requirements for reporting measure results.

The revised MSPB Hospital measure's use in CMS programs, like the HVBP program, is expected after legislative public reporting requirements for the Hospital Inpatient Quality Reporting and HVBP program are met. The current MSPB Hospital measure is publicly reported and used in HVBP payment determination for measure scores derived from at least 25 episodes and analysis of the revised MSBP Hospital measure indicates that the 25-episode case minimum for public reporting can remain unchanged.

The previously endorsed MSPB Hospital measure is publicly reported on Hospital Compare and used in the HVBP Program for eligible hospitals that have at least 25 episodes.

S.13.5. Define benchmarking and comparative estimates

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

The MSPB Hospital measure can be scored against benchmarks for the purpose of inclusion in incentive payment or other performance measurement programs. In this way, value in healthcare can be recognized and incentivized. The Hospital VBP Program provides financial incentives to short-term acute hospitals based on their performance on selected quality measures. By measuring the cost of care through the MSPB Hospital measure, CMS aims to recognize hospitals that can provide high quality care at a lower cost to Medicare. Combined with the other quality measures that comprise the Total Performance Score (TPS) under the Hospital VBP Program, the MSPB Hospital measure allows CMS to assess the value of care and incentivize both achievement and improvement in efficiency.

Under the Hospital VBP Program, hospital performance on the MSPB Hospital measure will be determined using the higher of its achievement or improvement score, as described in the FY 2012 IPPS Final Rule at 76 FR 51654-56. The MSPB Hospital measure score will then be included in the hospital's Total Performance Score (TPS) within the Efficiency and Cost Reduction domain. For information on how the MSPB-Hospital measure score was incorporated into the Hospital VBP Program, please refer to the FY 2012 IPPS/LTCH PPS final rule: http://www.gpo.gov/fdsys/pkg/FR-2011-08-18/pdf/2011-19719.pdf.

Validity – See attached Measure Testing Submission Form

SA.1. Attach measure testing form

2020-07-31-nqf-testing-form-mspb-hospital-v6-637318175300838758.docx

Measure Number (*if previously endorsed*)**:** 2158
Measure Title: Medicare Spending Per Beneficiary (MSPB) Hospital
Date of Submission: **8/3/2020**

Type of Measure:

Measure	Measure (continued)
☐ Outcome (*including PRO-PM*)	☐ Composite – *STOP – use composite testing form*
☐ Intermediate Clinical Outcome	☒ Cost/resource
☐ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	*

*cell intentionally left blank

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Long-term Minimum Data Set, Enrollment Database, and Common Medicare Environment - the Long-term minimum data set is used to obtain a single long-term care indicator for risk adjustment.	☒ other: Long-term Minimum Data Set, Enrollment Database (EDB), Common Medicare Environment (CME), American Community Survey (ACS), and Area Deprivation Index (ADI) – the ACS and ADI data are used specifically and only for social risk factor testing and elements from these data are ultimately not included in the measure specification.

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

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

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

Previous Response (2016): Medicare Parts A and B claims data from the Common Working File (CWF), Long-term Minimum Data Set (MDS) data, Enrollment Database (EDB) data, and the United States Census Bureau's American Community Survey.
Previous Response (2013): Medicare Parts A and B claims data from the Common Working File (CWF).

1.3. What are the dates of the data used in testing?

MSPB Hospital episodes from performance period 2018 (episodes with a discharge date occurring between January 1, 2018 and December 1, 2018) are used for almost all testing. Episodes from performance period 2017 (episodes with a discharge date occurring between January 1, 2017 and December 1, 2017) are included for select cross-year reliability testing. Please see Section 1.7 for more information on the data used in testing.

Previous Response (2016): Inpatient admissions with a discharge date between January 1, 2015 and December 1, 2015. For the test-retest analysis, data also included inpatient admissions with a discharge date between January 1, 2014 and December 1, 2014.

Previous Response (2013): May 15, 2010 – February 14, 2011

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

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

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

There are 3,218 acute care hospital providers with a MSPB Hospital measure score in the 2018 performance year, that are paid under Medicare's Inpatient Prospective Payment System, and that are located in 49 states (Maryland is excluded per CMS program requirements and reimbursement rates) and D.C. unless otherwise indicated. For most testing in this form, unless otherwise indicated, this sample of providers is limited to the 3,148 acute care hospital providers that also meet the 25-episode case minimum currently imposed on the MSPB Hospital measure under CMS programs and that testing in Section 2a2 indicates is still a high-reliability episode threshold for this measure.

Previous Response (2016): 3,298 Inpatient Prospective Payment System (IPPS) hospitals with discharges between 1/1/2015 and 12/1/2015 received an MSPB-Hospital measure value. Only claims for beneficiaries admitted to subsection (d) hospitals during the period of performance are included in the calculation of the MSPB-Hospital measure. Subsection (d) hospitals are hospitals in the 50 States and D.C. other than: psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer.

Previous response (2013): 3,396 IPPS hospitals received an MSPB Measure value (5/15/2010-2/14/2011 period of performance)

1.6. How many and which *patients* were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
There were 4,090,415 Medicare beneficiaries from 6,086,928 MSPB Hospital episodes included in hospital provider measure testing. These episodes span the measure's 2018 performance period. Generally, the beneficiaries included in the MSPB Hospital measure calculation are enrolled in Medicare Parts A and B (but not Part C) and have had an admission to an acute care hospital. Specifically, beneficiary episodes were included in the study sample if they met the following criteria.

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

To determine whether the MSPB Hospital measure's inclusion/exclusion criteria distort patient or episode characteristics, we analyzed distributions of patient characteristics (age, race, sex, dual eligibility status, hierarchical condition categories [HCCs]) and patient regional characteristics (e.g., income, unemployment) for (i) episodes with inclusion criteria, (ii) episodes without inclusion criteria, (iii) beneficiaries with inclusion criteria, and (iv) beneficiaries without inclusion criteria. The analysis demonstrated that the MSPB Hospital measure's inclusion criteria have a minimal effect on the percentage of beneficiary episodes defined by any

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

demographic (Appendix Table 1.6). For example, the percentage point difference for each demographic characteristic, before and after exclusion criteria application, ranged between -1.7 and +1.7 for episodes and between +1.4 and -1.4 for beneficiaries. The largest percentage point change from applying inclusion/exclusion criteria occurred in the study population's gender, as the proportion of female beneficiary episodes increased from 53.3 percent to 55.0 percent (episodes) and 54.1 percent to 55.5 percent (beneficiaries). Remaining differences in study characteristics were largely less than 1 percentage point after application of inclusion criteria. Section 2b2 discusses cost characteristics of the included/excluded populations.

Previous Response (2016): 4,261,069 beneficiaries (from 5,531,258 episodes) were included in the testing and analysis. These beneficiaries are enrolled in Medicare fee-for-service and were discharged from short-term acute hospitals between 1/1/2015 and 12/1/2015. Specifically, Medicare Part A and Medicare Part B claims from beneficiaries with an index admission within a subsection (d) hospital are included in the MSPB-Hospital episode if the beneficiary has been enrolled in Medicare Part A and Part B for the period 90 days prior to the start of an episode (i.e., 93 days prior to the date of the index admission) until 30 days after discharge.

To determine whether the MSPB-Hospital measure inclusion criteria distort patient characteristics on index admissions, we produced and analyzed distributions of patient characteristics (age, race, and sex) for two groups of patients: one group in which the beneficiaries had an eligible admission, and the other group in which patients both had an eligible admission and met the specified inclusion criteria as specified above. Appendix Tables 1-1, 1-2, and 1-3 detail these distributions and show that the MSPB-Hospital measure inclusion criteria do not significantly change the percentage of beneficiaries of any particular demographic. The typical difference between groups for a given characteristic is usually within 1 percentage point. To illustrate, the percent of beneficiaries aged 70 to 75 in the group that applies the inclusion criteria is 17%, compared to 16% when not implementing the inclusion criteria. The breakdown of race (i.e., Black and Non-Black) with and without the inclusion criteria is nearly identical. The breakdown of male and female beneficiaries with and without the inclusion criteria is also very similar, as the composition is 56% female in the group implementing the inclusion criteria compared to 55% when not applying the inclusion criteria.

Previous response (2013): 3,566,422 beneficiaries. These beneficiaries are enrolled Medicare fee-for-service and were discharged from short-term acute hospitals between (5/15/2010 and 2/14/2011)

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

MSPB Hospital episodes in the 2018 performance period include episodes whose triggering hospitalization (index admission) discharge date occurs from January 1 through December 1 of a calendar year and are not otherwise excluded by the criteria noted in Section 1.6. Social risk factor testing (Section 2b3) excludes approximately 1.6 percent of episodes in the 2018 performance period that cannot be matched to social risk factor data (e.g., ACS variables). Select reliability testing (Section 2a2) includes MSPB Hospital episodes from the 2017 performance year.

Previous Response (2016): N/A. The data samples used for the different aspects of testing below are identical. The test-retest analysis looked at data from one year prior as well, as noted in Section 1.3.

Previous response (2013): The data samples used for the different aspects of testing below are identical.

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

The social risk factors analyzed come from the ACS, ADI, EDB, and CME. All ACS variables are first defined at the Census Block Group level and then ZIP code when census block group is missing. The specific social risk factors (SRFs) analyzed include the following variables.

- Income (ACS): Low Income: median income < 33rd percentile nationally; Medium Income: median income in the interval spanning the 33rd percentile to the 66th percentile nationally; High Income: median income > 66th percentile
- Education (ACS): Education < High School: when % with < high school education is the highest for a given Census Block Group; Education = High School: when % with only high school is the highest; Education > High School: when % with > high school is the highest
- Employment (ACS): Unemployment Rate > 10%; Unemployment Rate <= 10%
- Race (EDB): Asian, Black, Hispanic, North American Native, White, and Other
- Sex (EDB): Female, male
- Dual status (CME): Full dual, partial dual, non-dual
- Area Deprivation Index (ADI)[2]: top quintile
- Agency of Healthcare Research and Quality (AHRQ) SES Index: AHRQ index scores are calculated using the AHRQ scoring algorithm and is a continuous dependent variable as a replacement of all SES variables. The index includes percentage of households containing one or more person per room, median value of owner-occupied dwelling, percentage of persons below the federally defined poverty line, median household income, percentage of persons aged ≥ 25 years with at least 4 years of college, percentage of persons aged ≥ 25 years with less than a 12th grade education, and percentage of persons aged 16 or older in the labor force who are unemployed.[3,4]

Previous Response (2016): The socioeconomic (SES) factor we analyzed is family income-to-poverty ratio. We obtained community-level poverty data from the 2014 American Community Survey, accessed through the United States Census Bureau's American FactFinder website, to determine the number of families in a given

[2] University of Wisconsin School of Medicine Public Health. 2015 Area Deprivation Index v2.0. Downloaded from https://www.neighborhoodatlas.medicine.wisc.edu/ February 24, 2020.

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

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

ZIP code whose income-to-poverty ratio (the ratio of family income to the federal poverty threshold) falls into certain categories. The dataset "Ratio of Income to Poverty Level of Families in the Past 12 Months" contains variables that represent ranges of income-to-poverty ratios. The values for these variables are the number of families in a given ZIP code whose income-to-poverty ratio falls into that variable's income-to-poverty ratio range. For example, if the value for the ".50 to .74" variable is 10,000 for a particular ZIP code, that means that 10,000 families in that ZIP code have incomes that are between 50% and 74% of the federal poverty threshold.

Enrollment Database (EDB) data provided the ZIP codes for beneficiaries included in the sample. We then linked these beneficiary ZIP codes to the ACS ZIP code-level data on family income-to-poverty ratio, which allowed us to analyze poverty data in beneficiaries' ZIP codes. We used family income-to-poverty ratio instead of individual income-to-poverty ratio to better reflect actual financial assets available to beneficiaries, as individual family members may pool financial resources to provide care for older relatives.

Previous Response (2013): n/a

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 used signal-to-noise and multi-sample analyses to test the reliability of the MSPB Hospital measure.

Signal-to-noise Analysis: Our signal-to-noise analysis sought to determine the extent to which variation in the measure is due to true, underlying provider performance, rather than variation within provider, from provider episodes. We calculated the reliability score for a hospital j as:

$$R_j = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_{w_j}^2}$$

Where $\sigma_{w_j}^2$ is the within-hospital variance of the mean measure score of hospital j, σ_b^2 is the between-hospital variance, and the measure's reliability score for hospital j, R_j, is calculated as the ratio of between-group variance to the sum of between-group variance and within-group variance. The closer a reliability score is to 1.0, the larger the between-group variance is relative to the within-group variance, the greater the suggestion that the measure is capturing the systematic differences between hospitals.

Multi-Sample Reliability Testing: Our multi-sample testing examined agreement between two hospital measure scores from (1) a randomly split set of episodes in the 2018 performance period and (2) the 2018 and 2017 performance periods. Only providers meeting an episode minimum of 25 episodes in studied samples were included. We analyzed score agreement from Pearson, Spearman, and Shrout-Fleiss intraclass correlation coefficients ICC(2,1). Coefficients close to 1.0 indicate high agreement in scoring between samples and suggest that performance scores are identified more by provider characteristics, like efficiency of care, than by random variation.

Previous Response (2016): *Data Element Reliability*: To construct the MSPB-Hospital measure, Acumen uses CMS claims data. CMS has in place several hospital auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures, including diagnosis and procedure codes and other elements that are consequential to payment. Specifically, CMS works with Program Safeguard Contractors (PSCs)/Zone Program Integrity Contractors (ZPICs) to ensure program integrity; the agency also uses Comprehensive Error Rate Testing (CERT) Contractors to ensure that Medicare payments are correct. Between 2005 and 2015, CERT estimates that proper payment, which is payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments each year.[5] CMS continues to perform successful corrective actions and give providers additional education to ensure accurate billing. To ensure claims completeness and inclusion of any corrections, the measure is calculated using data with a 3 month claims run-out from the end of the performance period.

Measure Reliability: Measure reliability is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. To estimate measure reliability, we utilize two approaches: (1) Test/Retest and (2) Reliability Score.

Our first approach to assess reliability is to consider the extent to which assessments of a hospital using unique sets of episodes produce similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured using two sets of episodes. We examine the correlation and quintile rank stability between a hospital's MSPB-Hospital scores calculated from both samples. By comparing the correlation of a hospital's MSPB measure calculated using the two mutually exclusive samples, one can identify the relationship of a hospital's score across samples. For this analysis, Acumen performed two separate test/retest investigations: comparing two random subsets of episodes from 2015, and comparing the set of 2015 episodes to the set of 2014 episodes. Both investigations sought to identify the reliability of a hospital's score across samples.

Our second approach calculates reliability scores as: $R_j = V_b/(V_b + (V_{w_j}/n_j))$ where R_j is the reliability for hospital *j*, V_b is the between hospital variance, V_{w_j} is the within hospital variance for hospital *j*, and n_j is the number of MSPB episodes for hospital *j*. This analysis seeks to determine the extent to which variation in the

[5] Comprehensive Error Rate Testing (CERT) Program. "Appendices Medicare Fee-for-Service 2015 Improper Payments Report". Table A6. https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicare-FFS-Compliance-Programs/CERT/CERT-Reports-Items/Downloads/AppendicesMedicareFee-for-Service2015ImproperPaymentsReport.pdf

measure is due to true, underlying hospital performance rather than random variation (i.e. statistical noise) within hospitals due to the sample of cases observed.

Previous response (2013): Data Element Reliability: Due to CMS's extensive auditing program, we believe that patient demographics, diagnostic information, and payment information are very reliable. As described in F.4., CMS uses various auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS also routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures.

Measure Reliability: The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is naturally the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. To estimate measure reliability, we utilize four approaches: (1) Test/Retest, (2) Seasonality, (3) Reliability Score, and (4) Bootstrapping.
Our first approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured once using a random subset of patients, then measured again using a second subset (over the same time period) that excludes the MSPB episodes chosen for the first sample. We examine the correlation, and quintile rank stability between a hospital's MSPB scores calculated from both samples.

Second, because the MSPB Measure values reported on Hospital Compare in April 2012 use Medicare claims data from May through February, Acumen conducted a seasonality analysis to examine how MS-DRGs change within a year. Providers that efficiently treat specific DRGs may receive higher MSPB Measure values during a season where the DRG occurs frequently and lower MSPB Measure values during a season where the DRG occurs less frequently. For this specific analysis, we split inpatient claims data with through date in 2010 into two categories: claims with through dates from January through April and claims with through dates from May through December.

Our third approach calculates reliability scores as: $R_j = V_b/(V_b + (V_{w_j}/n_j))$ where R_j is the reliability for Hospital j, V_b is the between hospital variance, V_{w_j} is the within hospital variance for hospital j, and n_j is the number of MSPB episodes for hospital j.

Fourth, Acumen measured how reliability varies based on the number of MSPB episodes a hospital is assigned. This fourth analysis is divided into two parts. The first evaluates how the number of MSPB episodes a hospital receives affects its 95 percent confidence interval. This analysis also informs how CMS should set the minimum number of episode required for public reporting purposes. When increasing the threshold for the minimum number of cases (or hereafter referred to as 'episode'), one decreases the likelihood an outlier episode[6]

[6] Statistical outlier episodes are excluded from the MSPB calculation to mitigate the effect of high-cost and low-cost outliers on each hospital's MSPB Measure. The MSPB Measure methodology uses "residuals" to define outlier episodes, where a residual equals the standardized episode spending minus the expected episode spending. High-cost outliers are defined as episodes whose residual falls above the 99th percentile of the residual cost distribution within any MS-DRG

materially affects a hospital's MSPB score, but also decreases the number of hospitals able to publicly report their MSPB Measure.

Whereas determining the number of hospitals that would be dropped when the minimum episode threshold increases is straight-forward, our second approach for measuring the effect of the minimum episode threshold on the MSPB confidence interval requires additional explanation. Typically, confidence intervals are constructed for commonly used quantities, such as the sample mean in which the distribution of the sample quantity is known, and can be used in the interval calculation. However, the MSPB score is a ratio of weighted means and does not have an easily identifiable statistic that corresponds to dispersion. Further, the MSPB score is not normally distributed, and typical measures of the dispersion of a distribution—such as the standard deviation—will not fully characterize the variation in the MSPB distribution.

In this analysis, Acumen instead uses a non-parametric bootstrap methodology to measure how the confidence interval of the MSPB score changes when the minimum episode threshold increases. This analysis measures the MSPB score for an 'average' hospital, where the 'average' hospital case is considered to be one whose MSPB episode distribution mimics that of the entire population of MSPB episodes. The bootstrap simulates the process of randomly drawing MSPB episodes from the population, and thus approximates the actual shape of the MSPB score distribution from which confidence intervals are determined. By repeatedly calculating an MSPB score for this simulated hospital under differing assumptions on the number of episodes observed, one can create a confidence interval for the MSPB score of this 'average' hospital.

To implement the bootstrap procedure, this analysis examines cases where the 'average' hospital has X episodes, where X = 1, 2, 3, 5, 10, 25, and 100. The five step methodology used to implement this analysis is as follows: (1) Draw 10,000 random samples (with replacement) each with X number of episodes from the original dataset containing MSPB episodes; (2) Calculate MSPB Amount for each sample; (3) Calculate MSPB Measure—normalization of the MSPB Amount—as the MSPB Amount for the hospital divided by the median MSPB Amount across all hospitals; (4) Calculate the 95 percent confidence interval using the 2.5th and 97.5th percentiles of the MSPB Measure distribution;[7] and (5) Divide the width of this confidence interval by the width of the confidence interval for X = 100 episodes.

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. The average reliability score of hospitals with at least 25 episodes was 0.92, with 99.0 percent of providers meeting or exceeding a 0.4 reliability score, a standard generally considered as the

admission category; similarly, low-cost outliers are defined as episodes whose residual falls below the 1st percentile of the residual cost distribution within any MS-DRG category. For additional details on the definition of statistical outliers for the MSPB Measure, see the response to Question 2a1.20 of this measure submission form.

[7] If a hospital has a true MSPB Measure value of 1.0, a 95% confidence interval indicates that 95% of the time the hospital's MSPB Measure value will fall between the 2.5th and 97.5th percentiles if the hospital gets X number of episodes from the original dataset containing MSPB episodes.

threshold for 'moderate' reliability[8], and 94.3 percent of providers meeting or exceeding a 0.7 reliability score (Appendix Table 2a23.a). While higher episode-minimums yield higher reliability results, the application of higher episode-minimums reduces the number of providers receiving a measure score. The median reliability score for hospitals with at least 25 episodes was 0.96 and the reliability score interquartile range spanned from 0.91 to 0.98 (Table 1).

Table 1. Distribution of Reliability Scores for Providers with at Least 25 Episodes

Number of Hospitals	Mean (Std. Dev.)	25th Pct.	50th Pct.	75th Pct.
3,148	0.92 (0.12)	0.91	0.96	0.98

* Pct. = percentile.

Split-sample Reliability Testing Results. The Pearson correlation coefficient was 0.83 for the 2018 split-sample and 0.79 for the 2017 and 2018 sample (Table 2, Appendix Table 2a23b). The Shrout-Fleiss intraclass correlation coefficients were similar at 0.83 and 0.79 for the 2018 split-sample and 2017 and 2018 sample.

Table 2. Split-sample and Two-Year Sample Correlation Coefficients for Hospitals with At Least 25 Episodes

Sample	Pearson Correlation Coefficient	ICC(2,1)
2018 Random Split	0.8265	0.8264
2018 and 2017 performance periods	0.7910	0.7873

Previous Response (2016):

1. Test/Re-Test: For the 2014 and 2015 sample (i.e., comparing 2015 data to 2014 data), over 75 percent of hospitals in the lowest-spending quintile in one year are in the lowest-spending quintile in the other; similarly, over 74 percent of hospitals in the highest-spending quintile in one year are in the highest-spending quintile in the other. Moreover, over 91 percent of hospitals in the highest-spending quintile in one year are in one of the top two highest spending quintiles in the other year. Quintiles results are listed in Appendix Table 2a2-1. The Spearman rank correlation for a hospital across the two years is 0.85, and the Pearson correlation coefficient is 0.81. As a point of comparison, in a standard moving-average time series process with one lag (i.e., an MA(1) process), the maximum possible Pearson correlation is 0.50.[9] Therefore, the value of 0.81 is remarkably high in relation to a relevant statistical benchmark. For the 2015 sample (i.e., comparing two random subsets of episodes from 2015), over 72 percent of hospitals in the lowest-spending quintile in one sample are in the lowest-spending quintile in the next; similarly, over 71 percent of hospitals in the highest-spending quintile in one sample are in the highest-spending quintile in the next. Moreover, over 90 percent of hospitals in

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

[9] Goldberger, 1991, A Course in Econometrics, and Greene, 2002, Econometric Analysis. An MA(1) model of a dependent variable such as the MSPB score takes the form $y_t = \mu + u_t + \theta u_{t-1}$, where t indicates the time period, μ is a constant over time, and u_t and u_{t-1} are mean zero, independent error terms.

the highest-spending quintile in one sample are in one of the top two highest spending quintiles in the next. The Spearman rank correlation for a hospital across samples is 0.82, and the Pearson correlation coefficient is 0.70. In a simple econometric model where two outcomes share a mean and each have two additive error terms (one in common, and one distinct), the Pearson correlation is 0.50. [10] The value of 0.70 is high relative to this statistical benchmark in which the expected value of the two outcomes are completely identical.

2. Reliability Score: Using a minimum episode threshold of 25 MSPB-Hospital episodes, over 99 percent of hospitals have a reliability score greater than 0.4 and 67.9 percent of hospitals have a reliability score greater than 0.9. Additionally, the average reliability score for hospitals with at least 25 episodes is 0.897. Previous work supported that 0.4 is the lower limit of "moderate" reliability;**Error! Bookmark not defined.** the MSPB-Hospital measure exceeds this threshold for over 99 percent of hospitals.

Previous Response (2016) Appendix A: The original MSPB-Hospital measure submission demonstrated measure score reliability using two analyses: calculation of measure reliability scores and a test-retest analysis. The measure reliability score calculation showed the percentage of hospitals with a reliability score greater than 0.4 and a reliability score greater than 0.9 for hospitals with at least 25 MSPB-hospital episodes, and the original test-retest analysis compared movement of hospital measure scores across quintiles.

NQF Committee Feedback: The NQF committee commented on two aspects of the original submission's reliability analyses. First, some committee members requested a more granular breakdown of reliability, citing a reliability threshold of 0.7 and asking about the effect of case minimums on measure reliability. Second, regarding the test-retest analysis, a committee member noted that approximately 30% of hospitals in the lowest spending quintile in one sample were not in the lowest spending quintile in the other sample.

Methods: To address the committee feedback, Acumen performed two additional analyses: (i) calculation of reliability numbers at additional thresholds, and (ii) an expansion of the test-retest analysis. For the reliability analysis, Acumen calculated reliability using the same methodology as the original submission. Supplementary Table 1 below shows the percentage of hospitals with reliability greater than the 0.4 and 0.7 thresholds for case minimums of 25 episodes, 40 episodes, 60 episodes, and 80 episodes. For the test-retest analysis, Acumen used the same methodology as in the original submission. However, the updated analysis shows movement of providers in the lowest 40th percentile of spending, rather than analyzing movement across quintiles.

Results: Supplementary Table 1 shows the percentage of providers with reliability greater than or equal to 0.4 and 0.7, for episode case minimums of 25, 40, 60, and 80. Of the 3,211 providers meeting the 25 episode case minimum, 99.1% have reliability greater than or equal to 0.4. This number is also high for reliability greater than or equal to 0.7, where 93.1% of providers meet the threshold. In addition, the percentage of providers meeting the reliability thresholds of 0.4 and 0.7 increases very little as the case minimum increases from 25.

[10] This example parallels the MA(1) time series example in footnote 2; see the references there for details. The econometric model of two outcomes in time period t, y_{t1} and y_{t2}, is given by $y_{t1} = \mu + \epsilon_t + u_{t1}$ and $y_{t2} = \mu + \epsilon_t + u_{t2}$, where μ is the shared mean, and ϵ_t, u_{t1} and u_{t2} are independent, mean zero error terms with common variance.

The test-retest analysis shows that, when comparing 2014 and 2015 data, 84% of providers in the lowest 40th percentile of spending for one sample are also in the lowest 40th percentile of spending for the other sample. Supplementary Table 2 shows full results for the test-retest analysis.

Supplementary Table 1: Provider Measure Reliability Breakdown

# of Threshold Episodes	% of Providers with Greater than or Equal to 0.4 Reliability	% of Providers with Greater than or Equal to 0.7 Reliability	# of Providers With This Many Episodes	% of Providers With This Many Episodes
Greater Than or Equal to 25 Episodes	99.1%	93.1%	3,211	97%
Greater Than or Equal to 40 Episodes	99.5%	93.3%	3,182	96%
Greater Than or Equal to 60 Episodes	99.6%	93.3%	3,144	95%
Greater Than or Equal to 80 Episodes	99.8%	93.5%	3,100	94%

Supplementary Table 2: Test-Retest Measure Score Movement

Method & Number of Episodes Restriction	# of Providers in Lowest 40th Percentile of Measures 1 & 2	# of Providers in Lowest 40th Percentile of Measure 1	# of Providers in Lowest 40th Percentile of Measure 2	% of Providers in Lowest 40th Percentile of Measure 1 that Are In Lowest 40th Percentile of Measures 1 & 2	% of Providers in Lowest 40th Percentile of Measure 2 that Are In Lowest 40th Percentile of Measures 1 & 2
All Providers in 2014 & 2015 Data	1,098	1,309	1,309	84%	84%
All Providers with At Least 25 Episodes in 2014 & 2015 Data	1,073	1,275	1,275	84%	84%

Previous Response (2013):

1. Test/Re-Test: Over 70 percent of hospitals in the lowest-spending quintile in one sample are in the lowest-spending quintile in the next; similarly, over 70 percent of hospitals in the highest-spending quintile in one sample are in the highest-spending quintile in the next. The Spearman rank correlation for a hospital across samples is 0.835.
2. Seasonality Analysis: Between the January 2010 – April 2010 period and the May 2010 – December 2010 period, the average absolute change in the relative frequency of an MS-DRG index admission was 8.9%. Certain lung-related admissions (e.g., pneumonia, COPD, asthma) appear more frequently in the winter.

3. Reliability Score: The MSPB Measure's overall reliability is 0.951. Over 98 percent of hospitals have a reliability score greater than 0.4; 62 percent of hospitals have a reliability score greater than 0.9. Previous work proposed that 0.4 is the lower limit of "moderate" reliability;[Error! Bookmark not defined.] the MSPB measure exceeds this threshold.
4. Minimum Number of Cases Required for the MSPB Measure: As the minimum episode threshold increases, there is a trade-off between the size of the confidence interval for the 'average' hospital and the number of hospitals receiving an MSPB score. Table 1 in the appendix shows that as the minimum episode threshold, X, increases, the confidence interval becomes narrower and more reliable. Specifically, the 95% confidence interval decreases by almost a third as cutoff number is moved from X = 5 to X = 50. However, as the minimum episode threshold increases from X = 5 to X = 50, the number of hospitals that could publicly report this measure included decreases; in fact, at the cutoff X = 50 episodes, the share of hospitals included decreases to 95.9%.

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, the reliability of the MSPB Hospital measure is high, including when its current 25-episode minimum is applied to balance measure reliability and inclusiveness.[11] The MSPB Hospital measure performance period episode minimum is 25 for the HVBP program, and the signal-to-noise analysis indicates that this episode minimum maintains the measure's high reliability.

The correlation coefficients for scores across the 2018 and 2017 performance periods were lower than scores compared across the randomly split 2018 performance period sample. This difference is expected as the two-year sample may capture additional variation in hospital performance across performance periods. The Shrout-Fleiss intraclass correlation coefficients were similar to the Pearson correlation coefficients at 0.83 and 0.79 for the 2018 split-sample and 2017 and 2018 sample. As ICC(2,1) imposes a common variance for provider across samples, its use is most appropriate in assessing the reliability of the 2018 performance period random split-sample.

Previous Response (2016):

1. *Test/Retest*: Sample selection does not have a material effect on a hospital's MSPB-Hospital measure for different data samples drawn from the same period, or for data samples drawn from different

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

periods. . In other words, hospitals have similar MSPB-Hospital measure quintile ranks regardless of which MSPB-Hospital episodes are used to calculate the MSPB-Hospital measure scores. This indicates that the MSPB-Hospital measure score is a reliable measure of a hospital's risk-adjusted Medicare spending compared to other hospitals.

2. *Reliability Score*: Overall reliability of the MSPB-Hospital measure is extremely high due to the large number of MSPB-Hospital episodes attributed to most hospitals. Reporting the MSPB-Hospital measure for hospitals that have at least 25 attributed episodes provides a balance between reliability and measure inclusiveness.

Previous Response (2016) Appendix A:

The updated reliability analysis shows that the overall reliability of the MSPB-Hospital measure is high, with roughly 93% of hospitals meeting the 0.7 reliability threshold even at the lowest case minimum. The 0.7 reliability threshold was mentioned by the NQF committee as an appropriate threshold for high reliability.

The MSPB-Hospital measure scores are stable across years when reviewing the hospitals in the lowest 40% of spending. Together with the original analysis demonstrating high correlation of the measure across samples and stability across quintiles for the large majority of hospitals, this further supports measure reliability.

Previous Response (2013):

1. Quintile Rank Stability Across Groups: Sample selection does not have a material effect on a hospital's MSPB score for different data samples drawn from the same period.
2. Seasonality Analysis: The seasonality analysis indicates that the incidence of different types of hospitalizations (i.e., MS-DRGs) varies across the year, but this variability for the most part is concentrated in DRGs lung-related diseases.
3. Reliability Score: Overall reliability of the MSPB score is extremely high due to the large number of MSPB episodes attributed to most hospitals. Reporting the MSPB Measure for hospitals that have at least 25 attributed episodes provides a balance between reliability and measure inclusiveness.
4. Minimum Number of Cases Required for the MSPB Measure: Based on the empirical results presented in 2a2.3., reporting the MSPB Measure as part of the Hospital VBP program for hospitals that have at least 25 attributed episodes provides a balance between the size of the confidence interval and the number of hospitals receiving and MSPB Measure score.

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish*

good from poor performance) NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

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

The MSPB Hospital measure went through the FY2012 rule-making cycle, receiving comment from public stakeholders, and was finalized in the FY2012 Final Rule. The MSPB Hospital measure was also endorsed by NQF, including across validity and reliability dimensions, in 2013 and 2016. In this section, we provide updated validity testing for the refined MSPB Hospital measure and detail on refinements that were considered.

Face validity

Potential refinements to the MSPB Hospital measure methodology that is in current use were identified from prior rule comments, past NQF endorsement cycles, and related measure development (e.g., MSPB Clinician). These potential refinements were tested and reviewed by a technical expert panel in February 2020 as part of the MSPB Hospital measure's re-evaluation. The TEP comprised 20 members with expertise in cost measure development and evaluation and quality improvement from diverse backgrounds, including clinicians, healthcare providers, academia, and patient advocacy organizations. Though no official vote was taken, panelists agreed that maintaining MSPB Hospital measure's holistic "all-cost" approach, allowing readmissions to trigger new MSPB Hospital episodes to increase measure surveillance, and updating the MSPB Hospital measure's MSPB Amount (score numerator) calculation to evenly weight all of a hospital's episodes were appropriate refinements. Panelists further provided additional considerations for ongoing social risk factor testing, like examining the impact of controlling for the Area Deprivation Index (Section 2b3.4.b details SRF **testing).**

Empirical Validity Testing

We undertook three approaches to empirically examine the extent to which the MSPB Hospital measure captures what it intends to capture. First, we examined the relationship between risk-adjusted episode cost ratios and episodes with and without post-admission events that are known indicators of high cost or intensive care. Specifically, we examined the observed to expected cost (O/E) ratios of episodes with acute care readmissions, episodes with any post-acute care (PAC) facility use, and episodes with PAC skilled nursing facility (SNF) use. We examined episodes with PAC-SNF use separately as such use has traditionally accounted for the largest share of Medicare's fee-for-service PAC expenditures.[12] As these post-index admission events are not directly controlled for through risk adjustment (although they are indirectly controlled for by the clinical risk adjustors such as MS-DRGs and LTI indicator), we would expect episodes that have such events will evidence observed episode costs that are higher than the cost predicted by risk adjustment – that is, we would expect O/E cost ratios for these episodes to be greater than 1.0 to the extent that the use of such post-admission services was not associated with clinical factors in the measure's risk adjustment model (e.g., other patient and provider considerations). Further, we would expect their counterpart episodes – episodes without such events – to have O/E cost ratios less than 1.0.

[12] http://www.medpac.gov/docs/default-source/data-book/jun19_databook_sec8_sec.pdf?sfvrsn=0

Second, we examined the relationship between a hospital's average expected episode cost (the average "E" in O/E cost ratios) and average episode rates of several service use categories. Per episode service use, particularly for higher cost events or events that require further care, like surgical procedures, may be positively correlated with expected episode costs if the regression model that the MSPB Hospital measure uses for risk adjustment predicts patient need for such services well. Section 2b3.3a. discusses how the MSPB Hospital measure's risk adjustment regression model, which is broadly based on the CMS HCC model, meets this prediction need. While we acknowledge that the hypothesized positive relationship between a hospital's average predicted episode cost and average episode rates of service use may not be linear or strong as high service use may be comprised of low-cost services relative to higher cost alternative services,[13] we would expect at least weakly positive rank relationships between a hospital's average expected episode cost and average per episode service use.

Finally, we examined the relationship between the MSPB Hospital measure and other cost-specific measures, efficiency-related measures, and measures in other HVBP program domains.[14] Any relationship between the MSPB Hospital measure and other measures may be obscured by many factors, including different measurement periods, populations, risk adjustment methods, or scoring methodologies. For example, while a MSPB Hospital measure performance period includes episodes from a single calendar year, measures in the HVBP program's Clinical Outcome domain rely on a performance period that spans 4 years. Further, while the MSPB Hospital measure for a hospital is scored relative to the episode-weighted median hospital's risk-adjusted cost, other measures in current use are scored relative to the mean hospital performance or relative to the total number of survey questions answered.

Thus, in this final analysis, we sought to compare MSPB Hospital measure components that may more closely relate to other measure scores and rates. Specifically, we compared the average expected episode amount to other measure performance period rates, for measures that had a literature-based or hypothesized conceptual relationship to the MSPB Hospital measure. We would expect the hospitals' average expected episode cost to be positively correlated with another cost-specific measure if the other measure's population is significant in terms of size or average costliness. Based on these characteristics, we examined the relationship between the MSPB Hospital measure's average expected episode cost and condition-specific Medicare cost measures that are also defined by inpatient hospitalization. We would also expect hospitals' average expected episode cost to be positively correlated with non-cost hospital measures that might speak to broader hospital efficiency. To test this expectation, we examined the relationship between a hospital's averaged expected episode cost and emergency department wait times that patients' face. Third, we would expect measures in other HVBP domains to relate to the MSPB Hospital measure's average expected episode cost positively in as much as measures in these other domains imply inefficiency or an excess of resources provided.

Previous Response (2016): Acumen utilized three tests to evaluate the validity of the MSPB-Hospital measure: (1) correlation with another measure of Medicare spending, specifically CMS' measure of risk-adjusted, standardized total Medicare spending at the Hospital Referral Regions (HRR) level, (2) correlation with service

[13] Consider, for example, the substitution between a high E&M visits per episode rate for regular patient check-ups versus a low but costly adverse event, like emergency surgery.

[14] The MSPB Hospital measure is used in one of four Hospital Value-Based Purchasing Program domains, the Cost and Efficiency Domain.

utilization rates, and (3) cost variation by time period. The first two correlations seek to confirm the validity of the MSPB-Hospital measure by comparing it with other measures of resource use, while the third test seeks to confirm the measure's validity by determining if cost variation by time period is consistent with expectations.

The first test examined the correlation between the MSPB-Hospital measure and the measure of risk-adjusted, aggregated annual per-capita spending for all Medicare beneficiaries produced by CMS at the HRR level.[15] This measure included all Medicare beneficiaries that had no months of Medicare Advantage enrollment and had both Part A and Part B for the portion of the year that they were covered by Medicare. Data on this measure of Medicare spending were available for 2007 – 2014, and Acumen performed correlation analyses for each of those years. For each HRR, Acumen found the mean MSPB-Hospital measure and correlated with the risk-adjusted, standardized, per capita HRR-level measure of total Medicare spending. This analysis sought to confirm the accuracy of the MSPB-Hospital measure by comparing its findings to a measure of Medicare spending.

The second test examined the correlation between the MSPB-Hospital measure and a measure of service utilization constructed by Acumen. To construct the service utilization measure, Acumen constructed hospital-level averages of services billed during the MSPB-Hospital episode across various categories (professional Evaluation & Management (E&M), post-acute, etc.). Acumen subsequently correlated these averages with the MSPB-Hospital measure. This analysis sought to confirm the expectation that the MSPB-Hospital measure correlates with service utilization rates.

The third test examined cost variation by time period. To do so, we broke down the total variance in risk-adjusted cost by time period, namely the period 3 days prior to and during the index admission and the period post-discharge. Because the risk adjustment model controls for MS-DRG, and because the MS-DRG of the index admission is the primary driver of costs from 3 days prior and during the index admission, the expected result of this analysis is that risk-adjusted episode cost should be strongly driven by post-discharge cost.

Previous Response (2013): The first validity test examines the correlation between hospitals' MSPB scores and the percent of beneficiaries with multiple episodes. This analysis examines whether high-cost hospitals may have below average (i.e., efficient) MSPB Measure values if the MSPB episode definition separates a single episode of care into two or more MSPB episodes. Division of a single episode of care into multiple MSPB episodes occurs when a hospital admission takes place more than 30 days after the initial discharge.

The second test of the validity of the MSPB Measure compares the MSPB Measure against other related outcome measures. Specifically, we will examine whether hospitals with low MSPB scores (i.e., efficient hospitals) are also less likely to have various types of hospital readmissions.

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

[15] Centers for Medicare & Medicaid Services. "Medicare Geographic Variation Public Use File." http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Geographic-Variation/GV_PUF.html

Empirical Validity

Observed to Expected Cost Ratios. The mean, standard deviation, and percentile distribution of observed to expected episode cost ratios for episodes with high-cost post-admission events were higher than their counter parts (Table 3, Appendix Table 2b1.3a). For example, episodes with an acute care rehospitalization an average O/E ratio of 1.55 and an interquartile range of 1.07 to 1.85, while episodes without such readmissions had an average O/E ratio of 0.89 and an interquartile range of 0.60 to 1.02.

Table 3. Distribution of Observed to Expected Ratios

Observed to Expected Ratios: Cost Driver Category	Observed to Expected Ratios: Mean	Observed to Expected Ratios: Std. Dev.	Observed to Expected Ratios: Percentiles (10^{th})	Observed to Expected Ratios: Percentiles (25^{th})	Observed to Expected Ratios: Percentiles (50^{th})	Observed to Expected Ratios: Percentiles (75^{th})	Observed to Expected Ratios: Percentiles (90^{th})
All Final Episodes	1.00	0.56	0.51	0.63	0.84	1.20	1.72
Episodes with downstream acute (re)admission	1.55	0.69	0.88	1.07	1.38	1.85	2.44
Episodes without downstream acute (re)admission	0.89	0.45	0.49	0.60	0.77	1.02	1.46
Episodes with Post-Acute Care (IRF, LTCH, HH, SN)	1.25	0.60	0.66	0.83	1.10	1.52	2.03
Episodes without Post-Acute Care (IRF, LTCH, HH, SN)	0.78	0.41	0.46	0.55	0.69	0.88	1.18
Episodes with Post-Acute Care SNF	1.43	0.59	0.83	1.02	1.30	1.69	2.18
Episodes without Post-Acute Care SNF	0.86	0.47	0.48	0.58	0.74	0.96	1.39

Service Utilization. Most service use/setting categories were moderately and positively correlated to the average predicted episode cost, with the correlations across all services categories average +0.487 and procedure use evidencing the strongest correlation (+0.721; Appendix Table 2b1.3b).

Other Measures. All three Payment & Value of Care measures, capturing 30-day Medicare payments for acute myocardial infarction, heart failure, and pneumonia conditions, were positively and weakly (or moderately) correlated with the hospital average predicted episode cost (Table 4, Appendix Table 2b1.3c). All four Timely & Effective Care measures, capturing time spent in the ED before being sent home or admitted, were also positively and weakly or moderately correlated with average predicted episode costs.[16]

[16] Timely and Effective Care measures from Hospital Compare archived data also included a measure of the percentage of patients who had cataract surgery and had improvements in visual function within 90-days. This measure was excluded from analysis due to its lack of a conceptual basis for relationship with the MSPB Hospital measure and small matched sample size (N=45).

The interpretation of performance rates for measures included in HVBP program domains varies by measure. In some cases, low performance rates are more desirable while in others high performance rates are better. Lower performance rates are better for HVBP Safety domain measures, which include rates of several healthcare-associated infections like catheter-associated urinary tract infections and *Clostridium difficile* infection. Higher performance rates are better for HVBP Clinical Outcomes domain measures,[17] which include 30-day condition-specific mortality measures, as these measures are expressed in terms of survival rates. Higher performance rates are also better for HVBP Patient Care & Experience domain measures, which include several HCAHPS questions on patient perceptions on staff, nurse, and physician communication, facility cleanliness, and care transitions. The MSPB Hospital measure's average expected episode cost was positively and weakly correlated with HVBP Safety domain measures (higher expected episode costs were positively related to HAI rates), positively and weakly correlated with HVBP Clinical Outcome survival rate measures (higher expected episode costs were positively related to condition-specific survival rates), and largely negatively and weakly correlated with patient perceptions, from HCAHPS survey questions, on hospital staff communications, cleanliness, and care transition planning (higher expected episode costs were negatively related to patient perceptions of hospital communication and efficiency).

Table 4. Spearman Correlation Statistics between Hospital Average Predicted Episode Cost and Other Measure Performance Rates

Measure	Range of Spearman Correlation Coefficient
Payment & Value of Care (AMI, HF, PN measures)	+0.13 to +0.49
Timely and Effective Care: Average/Median Time Spent Before Being Sent Home or Admitted	+0.26 to +0.45
HVBP Clinical Outcome Domain Measures (AMI, HF, PN survival performance period rates)	+0.13 to +0.20
HVBP Patient Care and Experience Domain Measures (HCAHPS questions on communication cleanliness, and care transitions)	-0.38 to +0.04
HVBP Safety Domain Measures (HAI 01-06, PC01)	+0.06 to +0.17

Previous Response (2016)

Correlation with Another Measure of Medicare Spending: For each year for which the risk-adjusted, standardized, per capita HRR-level measure data were available (2007 to 2014), the MSPB-Hospital measure had a positive correlation of at least 0.5 with the corresponding HRR-level measure. From 2007 to 2014, the lowest Spearman rank correlation for a given year was 0.53 and the lowest Pearson correlation coefficient was 0.51; during the same period, the highest Spearman rank correlation was 0.63 and the highest Pearson correlation coefficient was 0.61.

Correlation with Service Utilization Rates: The MSPB-Hospital measure had a Pearson correlation of 0.42 with professional E&M services per episode and a Pearson correlation of 0.52 with post-acute skilled nursing and inpatient services per episode.

[17] The Hip/Knee complication measure is not included in this analysis.

Cost Variation by Time Period: For the MSPB-Hospital measure, costs during the post-discharge period account for over 84 percent of total MSPB-Hospital episode cost variance, while costs from the period 3 days prior to and during the index admission account for just over 11 percent of total episode cost variance. These results are also shown in Appendix Table 2b2-1.

Previous Response (2016) Appendix A: The MSPB-Hospital measure submission demonstrated measure validity using three analyses. The first analysis showed correlation of the MSPB-Hospital measure with a measure of per-capita spending at the Hospital Referral Region (HRR) level. The second analysis showed correlation with a measure of hospital-level averages of service utilization. Finally, the third analysis examined cost variation by time period in the MSPB-Hospital episode.

NQF Committee Feedback: A committee member noted that the original 2012 submission of the MSPB-Hospital method included correlation analyses with condition-specific readmission measures. The committee member asked why these analyses were not included in the current submission.

Methods: Acumen appreciates the committee's comment and looked into the impact of readmissions in general. To examine the effects of readmission, Acumen calculated expected cost for episodes with and without an inpatient (IP) hospital readmission. Specifically, the same MSPB-Hospital risk adjustment model was used to calculate expected cost with an additional flag included for whether an IP readmission occurred in the episode window. Acumen also calculated the measure score distribution for providers based on the percentage of a provider's episodes that included an IP readmission.

Analyses comparing the MSPB-Hospital measure with the condition-specific readmission measures were excluded in the 2016 submission because the condition-specific readmission measures examine hospital performance on a specific set of conditions, while the MSPB-Hospital measure is intended to capture hospital performance across all acute conditions. Consequently, comparisons could be misleading. Since MSPB-Hospital is an all cost measure that includes all conditions, Acumen thought it would be more appropriate to look at the correlation between MSPB-Hospital and another broad-based all cost measure (i.e., the HRR measure).

Results: The mean expected episode cost for episodes with an IP readmission was $24,144, while the mean expected episode cost for episodes without an IP readmission was $19,617. Supplementary Table 3 presents the mean expected cost for episodes with and without an IP readmission.

Supplementary Table 3: Expected Cost for Episodes with and without IP Readmission

Episode Includes IP Readmission	# of Episodes	Mean Expected Cost
No	4,366,851	$19,617
Yes	1,053,782	$24,144

Supplementary Table 4 presents the measure score distribution across providers with varying percentages of episodes that include an IP readmission.

Supplementary Table 4: MSPB-Hospital Measure Scores by % of IP Readmission Episodes

Range	# of Providers	Mean MSPB-Hospital Measure Score
1) 0% <= % of Episodes with Readmissions <= 5%	35	0.888
2) 5% < % of Episodes with Readmissions <= 10%	115	0.894
3) 10% < % of Episodes with Readmissions <= 15%	661	0.937
4) 15% < % of Episodes with Readmissions <= 20%	1301	0.975
5) 20% < % of Episodes with Readmissions <= 25%	768	1.015
6) 25% < % of Episodes with Readmissions <= 30%	251	1.070
7) 30% < % of Episodes with Readmissions <= 35%	59	1.133
8) 35% < % of Episodes with Readmissions <= 40%	15	1.212
9) 40% < % of Episodes with Readmissions	6	1.309

This table shows that the MSPB-Hospital measure score tends to increase as a provider's percentage of episodes that include an IP readmission increases.

Interpretation: The two analyses looking at readmissions show that IP readmissions correlate with higher episode cost. Episodes with IP readmissions have a higher expected cost for readmissions, and providers with more IP readmissions have the higher MSPB-Hospital scores on average. This supports the validity of the MSPB-Hospital measure, as it accurately captures the higher resource use associated with IP readmissions.

Previous response (2013):

1. *Beneficiaries with Multiple Episodes:* The analysis indicated a positive correlation between MSPB Measure values and the percent of beneficiaries with multiple episodes. The hospital-level correlation between the MSPB Measure and the percent of beneficiaries with multiple episodes was 0.13; when accounting for variation in the MS-DRG of the index admission when measuring readmission rates, the correlation between readmissions and the MSPB Measure increases slightly to 0.16.
2. *Correlation with Other Outcome Measures:* The MSPB Measure exhibits a positive correlation with a number of hospital readmission measures. The correlation between the MSPB Measure and Heart Attack, Heart Failure, and Pneumonia Readmission Rates are of 0.08, 0.07, and 0.06, respectively.

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?*)

As expected, the average O/E cost ratio for episodes with downstream events that are of high resource, like readmissions or PAC use, are higher than episodes without such events.

While we acknowledged in Section 2b1.2 that our hypothesized positive relationship between a hospital's average predicted episode cost and average episode rates of service use may not be linear or strong as high service use may be comprised of low-cost services relative to higher cost alternative services, the positive correlations evidenced are in line with our expectations.

The relationship between the MSPB Hospital measure's risk-adjusted episode cost and other cost, efficiency, outcome, and quality measures are largely in line with hypothesized and literature-based expectations. Like the MSPB Hospital measure, the three Payment & Value of Care measures analyzed are triggered by an index hospitalization and consider standardized amounts. Unlike the MSPB Hospital measure, the episode window for these measures run 30-days from hospitalization – instead of 30-days after hospital discharge and are specific to hospitalizations that have principal discharge diagnoses of AMI, HF, or PN. Importantly, these measures also prorate claim payments to their 30-day episode window and consider patient populations that expired, while the MSPB Hospital Measure does neither and these measures differ in their risk adjustment model methods.[18] With these differences, however, we capture an expected positive rank correlation with these condition-specific cost measures. Further, the positive rank correlation between a hospital's average expected episode cost and non-cost measures of inefficiency (e.g. ED wait time) is in-line with existing literature.[19]

The rank correlations with other measures used in the FY2019 HVBP program and the MSPB Hospital measure's average expected cost are also in line with expectations. Literature has found, for example, that hospital acquired infections are associated with higher Medicare costs[20] and this recognition is not new, with CMS ceasing payment for select HAIs in the past.[21] Other literature has also noted the positive relationship between reported patient satisfaction and efficiency outcomes, like shorter stays, lower readmissions, and lower mortality rates, that can influence cost.[22]

Previous Response (2016): The interpretation of correlation results can depend on the specific analysis. In a simple econometric model where two outcomes share a common mean with additive and identically

---

[18] QualityNet, Hospital - Inpatient, Payment Measure Methodology (https://www.qualitynet.org/inpatient/measures/payment/methodology)

[19] Kyriacou, D. N., Ricketts, V., Dyne, P. L., Mccollough, M. D., & Talan, D. A. (1999). A 5-Year Time Study Analysis of Emergency Department Patient Care Efficiency. Annals of Emergency Medicine, 34(3), 326-335. doi:10.1016/s0196-0644(99)70126-5

[20] Hassan, Mahmud, Howard P. Tuckman, Robert H. Patrick, David S. Kountz, and Jennifer L. Kohn. "Cost of Hospital-Acquired Infection." Hospital Topics 88, no. 3 (2010/08/31 2010): 82-89. https://doi.org/10.1080/00185868.2010.507124.

[21] Peasah SK, McKay NL, Harman JS, Al-Amin M, Cook RL. Medicare non-payment of hospital-acquired infections: infection rates three years post implementation. Medicare Medicaid Res Rev. 2013;3(3):mmrr.003.03.a08. Published 2013 Sep 25. doi:10.5600/mmrr.003.03.a08

[22] Tsai TC, Orav EJ, Jha AK. Patient satisfaction and quality of surgical care in US hospitals. Ann Surg. 2015;261(1):2-8. doi:10.1097/SLA.0000000000000765

distributed errors, the Pearson correlation is 0.5 (see previous footnotes in the reliability testing Section 2a2.3).[23]

1. *Correlation with Another Measure of Medicare Spending:* The positive correlation between the MSPB-Hospital measure and the risk-adjusted, standardized, per capita HRR-level measure of Medicare spending indicates that the MSPB-Hospital measure's identification of hospitals with high- or low risk-adjusted spending is consistent with a measure of Medicare spending.
2. *Correlation with Service Utilization Rates:* The positive correlation between the MSPB-Hospital measure and service utilization rates, specifically for E&M services and post-acute nursing and inpatient services, indicates that the MSPB-Hospital measure accurately captures higher resource use.
3. *Cost Variation by Time Period:* Variance in costs during the post-discharge period makes up a larger portion of total variance than variance in costs during the period 3 days prior to and during the index admission does. This finding is consistent with expectations. The risk adjustment model predicts a certain level of post-discharge spending based upon the beneficiary's prior health history and MS-DRG. This analysis shows that of the cost variance left over after this risk adjustment, most of it is driven by post-discharge spending. Variance in provider scores based on post-discharge spending emphasizes the importance of care transitions and care coordination in improving patient care.

Previous Response (2013):

1. Beneficiaries with Multiple Episodes: Hospitals are not likely to be postponing necessary re-admissions—and thus creating a new episode—to improve their MSPB Measure values. High-cost hospitals are not more likely to treat beneficiaries with multiple hospitalization episodes.
2. Correlation with Other Outcome Measures: The positive correlation between the MSPB Measure and Heart Attack, Heart Failure, and Pneumonia Readmission Rates indicate that hospitals that are more expensive generally have higher readmission rates. The correlation, however, is weak for all three readmission rates. A weak correlation can be explained by the fact that the MSPB Measure assesses the cost to Medicare of all services performed by hospitals and other healthcare providers during an MSPB episode. As a result, a hospital's MSPB Measure value is driven by both acute and post-acute spending.

2b2. EXCLUSIONS ANALYSIS

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

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

We can classify the MSPB Hospital measure's exclusions as exclusions imposed to promote episode and provider comparability and exclusions imposed as part of data processing and completeness. The first exclusion type includes the exclusion of

- Transfer- or death- related episodes
- Non-IPPS, Non-acute, or Critical access hospitals
- Inpatient facilities in excluded states and territories

[23] Goldberger, 1991, A Course in Econometrics, and Greene, 2002, Econometric Analysis.

Excluding episodes where the beneficiary died prior to the episode's 30-day post-discharge period or where attribution is muddled by transfer allows the measure to avoid censored episode windows and simplify provider attribution. Excluding non-IPPS hospitals, non-acute hospitals, critical access hospitals, and hospitals in excluded states and territories (e.g., Maryland, Guam) promotes service-cost comparability and general practice patterns (e.g., care at a long-term hospital versus care at an acute care hospital). Given the rationales for this first exclusion type, we expect excluded episodes to differ from non-excluded episodes in terms of their cost profiles and tested as much. Specifically, we compared distributions of observed costs and observed to expected (O/E) cost ratios for excluded episodes against non-excluded episodes. We calculated expected episode costs for excluded episodes by including these episodes in risk adjustment.

Examples of the second type of exclusion include excluding episodes that may have invalid or incomplete data – like a mortality event before admission or evidence of competing insurer payment that may mask service use. This second type of exclusion is discussed in Section 2b6 (Missing Data Analysis and Minimizing Bias) of this testing form.

Previous Response (2016): Acumen evaluated the validity of the measure exclusion criteria by producing impact analyses, which show the effect of recalculating the MSPB-Hospital measure while independently reversing each of the following exclusion criteria: (1) acute-to-acute transfer episodes;[24] (2) death episodes;[25] and (3) outlier episodes.[26] For (1), our analysis evaluated the impact of including transfer episodes on MSPB-Hospital measure scores. For (2), we re-calculated the MSPB-Hospital measure using beneficiaries who die during the episode. Specifically, we examined the percent of beneficiaries who die during the MSPB-Hospital episode and the effect that including death episodes had on hospital scores. For (3), we examined the effect of including outliers when calculating MSPB-Hospital measure scores instead of excluding outliers based on the distribution of residuals. Specifically, we examined the impact of top-coding episodes with risk-adjusted costs that are above the 99th percentile, where those episodes are assigned the cost of the episode at the 99th percentile. We also examined the impact of bottom-coding episodes with risk-adjusted costs that are below the 1st percentile, where those episodes are assigned the cost of the episode at the 1st percentile.

The measure also implements an exclusion criteria specific to inpatient admissions that are allowed to trigger a new MSPB-Hospital measure. Specifically, we do not allow inpatient admissions that occur within 30 days post-discharge of another inpatient admission to start a new MSPB-Hospital episode; we refer to this criteria as excluding overlapping episodes. For this exclusion (4), we analyzed the effect of including overlapping episodes when constructing the MSPB-Hospital episodes. To illustrate what this exclusion is, take an inpatient admission that triggers Episode A and see if the beneficiary has another inpatient admission within the 30-day post-discharge window of Episode A. If the beneficiary has a second qualifying admission within the 30-day

[24] Transfers, defined based on the claim discharge code, are not considered eligible as index admissions. In other words, these cases will not generate new MSPB-Hospital episodes; neither the hospital which transfers a patient to another short-term acute hospital nor the receiving short-term acute hospital will have an index admission attributed to them.

[25] Recall from S.9.1. that any episode where at any time during the episode the beneficiary dies is excluded from the MSPB-Hospital calculation.

[26] Recall from S.9.1. that MSPB-Hospital episodes whose relative scores fall above the 99th percentile or below the 1st percentile of the distribution of residuals are excluded from the MSPB-Hospital calculation.

post-discharge window of Episode A, do not allow the second admission to trigger Episode B. We evaluated the impact of this exclusion on MSPB-Hospital measures by re-calculating MSPB-Hospital with the previously-excluded episodes added back in, which was then compared to MSPB-Hospital measures calculated under the overlapping episodes exclusion.

Previous response (2013): Acumen evaluated the validity of the inclusion/exclusion criteria by producing impact analyses which show the effect of recalculating the MSPB Measure while independently reversing each of the following inclusion/exclusion criteria: (1) beneficiaries in Medicare Advantage; (2) beneficiaries in Medicare Part A only; (3) acute-to-acute transfers;[27] (4) death episodes;[28] and (5) outlier episodes.[29] With respect to (3), Acumen's analysis evaluates assigning transfers to the transferring hospital and to the receiving hospital. The first three restrictions occur because of incomplete data or problems attributing episodes to individual hospitals. For (4), we re-calculate the MSPB Measure using beneficiaries who die during the episode. Specifically, Acumen examined the percent of beneficiaries who die during the MSPB episode and after the MSPB episode and whether or not to calculate separate MSPB Measures for beneficiaries who died during the episode versus beneficiaries who did not die. For (5), we examine top-coding/bottom-coding distribution outliers in place of completely excluding them.
Acumen also conducted a number of analyses on potential exclusion criteria. These unimplemented exclusions include: (6) beneficiaries discharged against medical advice (AMA) and (7) dual-eligibles. Acumen's analysis evaluates not counting admissions in which the beneficiary was discharged AMA as an index admission. Although excluding patients discharged against medical advice would avoid attributing the costs of non-compliant beneficiaries to a hospital's MSPB Measure value, hospitals would be incentivized to encourage high-cost beneficiaries to leave against medical advice to avoid having their episode included in the hospital's MSPB Measure. We also evaluate (i) including a dual-eligible indicator in the MSPB risk-adjustment and (ii) examining MSPB scores separately for duals/non-duals.

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 5 presents the percentage of episodes captured by each exclusion, observed cost statistics, and observed over expected (O/E) cost ratios for the MSPB Hospital measure exclusions. Cost statistics are also

[27] Recall from S.9.1. that transfers, defined based on the claim discharge code, are not considered eligible as index admissions. In other words, these cases will not generate new MSPB episodes; neither the hospital which transfers a patient to another short-term acute hospital, nor the receiving short-term acute hospital will have an index admission attributed to them. The rationale for exclusion of these acute-to-acute transfer cases is that CMS wished to perform further analysis of hospital impacts and explore potential unintended consequences of attribution of the MSPB episode to either the transferring or the receiving hospital.

[28] Recall from S.9.1. that any episode where at any time during the episode the beneficiary becomes deceased is excluded from the MSPB calculation.

[29] Recall from S.9.1. that MSPB episodes whose relative scores fall above the 99th percentile or below the 1st percentile of the distribution of residuals (see 2a1.20 for a description of MSPB residuals) within each index admission MS-DRG are excluded from the MSPB calculation.

provided for the remaining set of episodes after the described exclusions are applied for comparison. Appendix Table 2b2.2 provides more detailed cost distributions for measure exclusions.

Table 5. Cost Statistics for Measure Exclusions

Episodes:	Episodes: #)	Episodes: %	Observed Cost: Mean	Observed Cost: Percentile (10^{th})	Observed Cost: Percentile (90^{th})	O/E: Mean	O/E: Percentile (10^{th})	O/E: Percentile (90^{th})
All Episodes Meeting Triggering Logic	9,662,702	100.00%	$24,662	$7,210	$47,439	1.00	0.47	1.77
Episodes in which Inpatient Stay had Transfers or Death Discharge Status Codes or episodes that overlapped with an IP Stay with Transfer or Death Discharge Status Codes	676,060	7.00%	$36,508	$10,766	$72,856	1.14	0.46	2.18
Episodes in which beneficiary Death occurred within 30 Days Post Discharge	881,953	9.13%	$26,522	$9,946	$49,330	0.94	0.49	1.67
Episodes in which Inpatient stay occurred in a non-Acute Hospital or in a Critical Access Care (CAH) hospital	1,105,999	11.45%	$30,589	$7,469	$59,207	1.14	0.46	2.02
Episodes with Inpatient Facility located in Excluded Regions	231,396	2.39%	$22,791	$6,639	$43,763	0.99	0.47	1.75
Remaining Episodes	6,086,932	62.99%	$23,499	$7,209	$44,864	1.00	0.50	1.72

Previous Response (2016):

Transfer Episodes: Episodes that include an acute-to-acute transfer account for 1.6% of total episodes. Episodes containing an acute-to-acute transfer have an average observed cost of $33,363 compared to an average expected cost of $21,068, resulting in an observed-to-expected cost ratio of 1.58. Episodes not containing an acute-to-acute transfer, on the other hand, have an average observed cost of $20,570 compared to an average expected cost of $20,774, resulting in a observed-to-expected cost ratio of 0.99 (Appendix Table 2b3-1). Rural hospitals tend to have a higher rate of transfers than urban hospitals (4.1% and 1.3%, respectively), so including transfer episodes that have higher observed-to-expected cost ratio in the MSPB-Hospital measure calculation would probably disproportionately worsen rural hospitals' scores. When including transfer episodes in the calculation of the MSPB-Hospital measure, 81% of hospitals' MSPB-Hospital

measure scores change by less than ±0.03, and less than 2% of hospitals' MSPB-Hospital measure scores change by more than ±0.10 (see Appendix Table 2b3-2 for full results). The correlation between MSPB-Hospital measure scores when excluding transfer episodes versus when including transfer episodes is 0.95.

Death Episodes: In approximately 8% of MSPB-Hospital episodes, the beneficiary dies before the end of the 30-day post-discharge period. Episodes in which the beneficiary dies during the episode window (denoted as "death episodes") appear more efficient than non-death episodes, as shown in Appendix Table 2b3-3. The average observed cost of death episodes is $21,041 compared to the expected cost of $24,980, resulting in an observed-to-expected cost ratio of 0.84. Comparatively, non-death episodes have an observed-to-expected cost ratio of 1.02 ($20,512 over $20,156). If death is included in measure calculation, 96% of hospitals' MSPB-Hospital measure scores change by less than ±0.03, and very few hospitals (less than 0.2%) see changes in MSPB-Hospital measure scores greater than ±0.10 (see Appendix Table 2b3-4). The correlation between MSPB-Hospital measure scores when excluding death episodes versus when allowing for inclusion of death episodes in measure calculation is 0.99.

Outlier Episodes: When including outlier episodes in measure calculation, about 2% of hospitals see an absolute change in their MSPB-Hospital measure score of greater than ±0.10, and 6% of hospitals' MSPB-Hospital measure scores change by greater than ±0.05. Appendix Table 2b3-5 further details the impact of including outliers on MSPB-Hospital measure scores. The correlation between MSPB-Hospital measure scores when excluding outliers versus when including outliers is 0.93.

Overlapping Episodes: Approximately 12% of episodes had their trigger inpatient admission within 30 days of the discharge date of the trigger inpatient admission of another episode (Appendix Table 2b3-6). If episodes with a trigger inpatient admission during the 30-day post-discharge period of another episode are included in MSPB-Hospital measure calculation, 97% of hospitals' MSPB-Hospital measure scores change by less than ±0.03, with a small proportion of hospitals (0.4%) experiencing changes in MSPB-Hospital measure scores greater than ±0.10 (see Appendix Table 2b3-7 for detailed results). The correlation of MSPB-Hospital measure scores before and after removing the overlapping episodes exclusion is 0.99.

Previous Response (2013):

Medicare Advantage or Part A Only: 25% of Medicare beneficiaries are enrolled in Medicare Advantage; about 10 percent of Medicare FFS beneficiaries are enrolled in Part A only.

Transfers: Episodes that include an acute-to-acute transfer account for 5% of total episodes. Episodes containing an acute-to-acute transfer have an average risk-adjusted spending of $25,151 per episode, while the average episode not containing an acute-to-acute transfer has an average risk-adjusted spending of $19,489 per episode. Because transfer episodes cost 29% more than non-transfer episodes on average, excluding transfer episodes eliminates a significant portion of MSPB episodes and Medicare payments. Small rural hospitals are the most likely facilities to transfer to large, urban hospitals (see Tables 2 and 3 in the appendix). Assigning transfer episodes to the transferring hospital has a larger effect on the MSPB Measure than assigning transfer episodes to the receiving hospital. When transfer episodes are assigned to the receiving hospital, 90% of hospitals experience a change in their MSPB Measure values of less than 3 percent, but only 80% of hospitals experience a change in their MSPB Measure values of less than 3 percent when transfer episodes are assigned to the transferring hospital (see Tables 4 and 5 in the appendix)

Death Episodes: In approximately 8.0% of MSPB episodes, the beneficiary dies before the end of the 30-day post-acute period. Death episodes are much more expensive than non-death episodes. Whereas death episodes cost $26,883 on average, non-death episodes cost $19,141, a 40% difference in average episode cost. Since death episodes are typically expensive, including death episodes in the MSPB Measure would increase the skewness of the episode cost distribution. Including death episodes (after outlier episodes have been excluded) increases the ratio of the 99th percentile cost to the median cost by 3 percent. If death is included as a variable in the 'risk-adjustment' model, death episodes are only 16 percent more expensive than non-death episodes.

Outlier Episodes: As an alternative to excluding outlier episodes from the MSPB Measure, outlier episodes can instead be top-coded and/or bottom-coded. Rather than excluding episodes that are outliers, top-coding/bottom-coding assigns outliers the value of an episode at a specified threshold. Tables 6 through 10 in the appendix present the impacts of top-coding/bottom-coding episodes at the 99.9th/0.1th, 99.5th/0.5th, 99.0th/1.0th, 98.0th/2.0th, and 95.0th/5.0th percentiles, respectively, compared to a baseline that excludes outlier episodes at the 99th and 1st percentiles of the risk-adjusted episode cost distribution. When top-coded/bottom-coded at the 99.9th/0.1th, 99.5th/0.5th, and 99.0th/1.0th percentiles, at least 85 percent of MSPB Measure values change less than 3 percent. However, when top-coded/bottom-coded at the 98.0th/2.0th, and 95.0th/5.0th percentiles, at least 95% of MSPB Measure values change less than 3 percent (see Table 11).

Discharged AMA: Not only do episodes with an AMA discharge code make up a small percent of MSPB episodes (0.7%), AMA episodes have lower risk-adjusted spending than non-AMA episodes. ($13,851 vs. $19,025 for non-AMA). About 99% of hospitals experienced a change in their MSPB Measure values less than one percentage point when excluding AMA episodes (see Table 12).

Dual-Eligibles: 30% of episodes are flagged as dual-eligible beneficiaries; 18% of hospitals assigned an MSPB Measure have a beneficiary population consisting of at least 50% dual-eligible beneficiaries. Dual-eligible beneficiaries have $859 extra spending per episode than non-dual-eligible beneficiaries. If dual eligible are excluded, 43% of hospitals experience a change in their MSPB value of more than 1 percentage point (Table 13); including dual eligible in the risk adjustment model increases the R2 of the model by less than 0.001 and causes 12% of hospitals to change their MSPB Measure by more than 1 percentage point (Table 14).

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*)

Some excluded populations, like facilities located in regions excluded by HVBP program design, did not differ greatly from non-excluded episodes in observed cost and O/E cost ratios (e.g., O/E: 0.99 for hospitals in excluded regions vs 1.00 for non-excluded episodes). To calculate their O/E cost ratios, these excluded populations were included (in a one-off manner) with the non-excluded ("Remaining Episodes") population's in risk adjustment. Thus, their relatively close O/E indicates that the MSPB Hospital's risk adjustment model performed well for these hospitals. This is expected, given the rationale we provide on the measure's risk adjustment model's ability to predict expense throughout Section 2b3.

Other excluded populations, like those where index admissions were paid differently than acute care IPPS hospitals (e.g., critical access hospitals), where episodes windows were truncated due to death, or where episode attribution was complicated by transfers did differ from non-excluded episodes in their observed episode cost and O/E cost ratio distributions. For example, transfer-related episodes and non-acute or critical access hospitals averaged an O/E cost ratio of approximately 1.14 to the non-excluded episode ratio of 1.00. Further, the distributions for these two exclusions were generally higher than that of non-excluded episodes.

Previous Response (2016):

Transfer Episodes: Because transfer episodes are more inefficient than non-transfer episodes, regardless of the type of hospital (urban or rural), there are two main problems with including transfer episodes. First, because the observed cost relative to the predicted cost is high for transfer episodes (partly due to partial or full payments for two inpatient stays), including transfer episodes in the MSPB-Hospital measure may likely increase the MSPB-Hospital measure score of those hospitals most often engaging in transfers. These hospitals may not always have the capacity to handle these cases, and CMS may have an interest in ensuring medically appropriate transfers occur. Second, excluding transfer episodes addresses stakeholder concerns that neither the admitting nor receiving hospital is fully able to coordinate care. Stakeholders find it inappropriate to hold the transferring hospital responsible for services rendered by the receiving hospital, and it also may not be appropriate to hold the receiving hospital responsible for issues that arose prior to admission of a transferred patient. As a result, transfer episodes are excluded from the MSPB-Hospital measure calculation.

Death Episodes: Cases where the beneficiary dies during the episode are not eligible to be included in the MSPB-Hospital measure. Though the difference between cost for death and non-death episodes is relatively small compared to other exclusions, there are a few explanations for the exclusion of death episodes. First, including death episodes in MSPB-Hospital measure calculation may create problematic incentives. Death episodes appear more efficient than non-death episodes; unlike non-death episodes, which have a slightly greater observed cost than expected cost, the observed cost for death episodes is much less than the expected cost. This is because beneficiaries with death episodes likely have shorter episodes (and therefore fewer services) than beneficiaries with non-death episodes with the same DRG. Because of this, including death episodes in MSPB-Hospital measure calculation may incentive low-quality care, as increased mortality rates could potentially improve hospitals' MSPB-Hospital measure scores by including episodes that appear more efficient. Second, episodes during which a beneficiary dies are "truncated;" in other words, costs that might have occurred if the beneficiary had not died are not observed due to death. Death episodes are incomplete episodes where significant data could be missing when death occurs early in the episode. To avoid including episodes of care with incomplete costs and problematic incentives, episodes during which a beneficiary dies are excluded from the MSPB-Hospital measure calculation.

Outlier Episodes: Outliers are excluded from the MSPB-Hospital measure calculation to avoid cases where a handful of high-cost and low-cost outliers have a disproportionate effect on each hospital's MSPB-Hospital measure score. While the correlation between the measure when excluding outliers versus when including outliers is extremely high (0.93), outlier episodes impact a small percentage of hospitals' MSPB-Hospital measure scores in a large and important way, as demonstrated by the differences in scores described in Appendix Table 2b3-5. The distribution of hospital risk-adjusted episode spending is significantly right-skewed: the 99th percentile is 3.6 times the value of the median, while the 1st percentile is less than half the value of the median. Excluding outliers based on risk-adjusted cost eliminates the episodes that deviate most from the spending levels one would have expected based on patient demographics and severity of illness.

Overlapping episodes: Episodes that begin during a prior episode's 30-day post-discharge period are excluded from MSPB-Hospital measure calculation. The impact of the exclusion on hospitals' MSPB-Hospital measure scores is minimal, and the correlation of the MSPB-Hospital measure calculated with and without implementing the overlapping episodes exclusion is high.

Previous response (2013):
Medicare Advantage or Part A Only: Due to missing claims problems, only beneficiaries enrolled in Medicare Parts A and B Fee-for-service are included in the sample.

Transfers: Adding transfers to the MSPB measure would significantly change hospital MSPB scores and make episode attribution more complicated. Assigning transfer episodes to the transferring hospital would avoid giving providers an incentive to transfer high-cost patients to game the system; however, once the transferring hospital transfers the patient, they may have little opportunity to coordinate or affect the patient's post-discharge care. Small rural hospitals, for example, often transfer patients in cases where they do not have the capacity to treat the patient within their current facilities. Assigning transfer episodes to the receiving hospital, however, incentivizes the initial hospital to transfer complex patients to improve their MSPB score. Further, post-acute care coordination may be difficult if the receiving hospital is out of area.[30] Public comment in the FY 2012 IPPS notice of proposed rulemaking voiced concern over attribution in transfer cases. In response, CMS excluded these types of transfers from the finalized MSPB Measure (76 FR 51621).

Death Episodes: In the baseline specification, cases where the beneficiary dies during the episode are not eligible to be included in the MSPB Measure. Episodes during which a beneficiary dies are "truncated"; in other words, costs that might have occurred if the beneficiary had not died are not observed due to death. To avoid including episodes of care with incomplete costs, episodes during which a beneficiary dies are excluded from the MSPB Measure calculation. As shown in 2b3.3., these episodes are typically high cost. In fact, the Dartmouth Atlas also notes that patients with chronic illness in their last two years of life account for about 32% of total Medicare spending, much of it going toward physician and hospital fees associated with repeated hospitalizations.[31] This evidence indicates that including death as a risk adjuster reduces the disparity in death/non-death episode cost. However, if death is a risk adjuster, hospitals could improve their MSPB score by increasing mortality rates. Further, using death as a risk adjuster implies that the risk adjustment model is no longer prospective, since events that occur during an episode now influence the model's expected cost.

[30] As an alternative to completely assigning transfer episodes to either the transferring hospital or the receiving hospital, transfer episode costs could be split between both hospitals. A simple 50/50 weighting scheme would be one potential solution. To implement a 50/50 weighting scheme, each hospital receives 50% of the observed cost in the MSPB Amount numerator and 50% of the expected in the denominator of the MSPB Amount risk-adjustment factor (αj). This weighting scheme, however, does not take into account the length of stay at each hospital or the fact that the receiving hospital is in control of post-discharge spending. More complicated alternative weighting schemes (e.g., assigning a fixed weight to the receiving hospital and splitting the remaining weight based on the relative number of days the patient spends at each hospital) could be tailored to the particular application of the MSPB Measure, but these approaches would also increase the complexity of the MSPB Measure methodology.

[31] http://www.dartmouthatlas.org/keyissues/issue.aspx?con=2944

Outlier Episodes: Outliers are excluded from the MSPB Measure calculation to avoid cases where a handful of high-cost and low-cost outliers have a disproportionate effect on each hospital's MSPB Measure score. The distribution of hospital risk-adjusted episode spending is significantly right-skewed: the 99th percentile is almost 4.5 times the value of the median, while the 1st percentile is only approximately 1/2 the value of the median. Excluding outliers based on risk-adjusted cost eliminates the episodes that deviate most from the spending levels one would expect based on patient demographics and severity of illness. Outliers are identified across all episodes rather than within a hospital; thus, some hospitals may have no outlier episodes excluded and others many have many.

Discharged AMA: Episodes with AMA index admissions should be eligible to be considered as index admissions, as the effect of excluding AMA episodes from the MSPB Measure calculation is minimal (as shown in Table 12). Additionally, episodes with an AMA discharge code make up a small percent of MSPB episodes, and AMA episodes on average have lower risk-adjusted spending than non-AMA episodes.

Dual-Eligibles: Medicare beneficiaries who are dually-eligible for Medicare and Medicaid are not excluded from the MSPB Measure to be consistent with NQF's position on not adjusting for potential demographic (sex or race) or socioeconomic factors.

_____

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

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

☐ **No risk adjustment or stratification**

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

☒ **Stratification by 26 risk categories**

☐ **Other,**

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

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

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

In addition, the risk adjustment model includes status indicator variables for whether the beneficiary qualifies for Medicare through Disability or has ESRD. The model also includes an indicator of whether the beneficiary

recently required long-term care. Beneficiaries who need to reside in long-term care facilities typically require more intensive care than beneficiaries who live in the community. These enrollment and long-term care status variables are non-diagnostic based indicators of severity of illness.
The model also accounts for disease interactions between HCCs and/or enrollment status variables included in the MA model. These interactions are included because certain combinations of comorbidities increase costs more than is predicted by the HCC indicators alone. The risk adjustment model now also includes an indicator for whether an episode's index admission was triggered within the 30 day post discharge period of another inpatient stay – to better predict the higher cost of readmission stays (Section 2b3a.3a provides more detail).

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

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

Appendix Table 2b3.6.b provides regression coefficients, standard errors and other statistics for each model.

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

2b3.3a. Describe the conceptual/clinical *and* statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?
Clinical Factors: The CMS-HCC model was selected based on previous studies evaluating its appropriateness for use in risk adjusting Medicare claims data. This model was developed specifically for use in the Medicare population, meaning that it accounts for conditions found in the Medicare population and is calibrated on Medicare fee-for-service beneficiaries. In addition, the CMS-HCC model is routinely updated for changes in coding practices (e.g., the transition from ICD-9 to ICD-10 codes) and is exhaustive on these code sets. Because

the CMS-HCC model has already been extensively tested and is used for a large Medicare Part C population, we focus our testing on how the CMS-HCC model was adapted to the MSPB Hospital measure.[32,33,34]

The statistical risk model is estimated separately for each MDC, which is determined by the MS-DRG of the index admission; in turn, these are generally grouped according to principal diagnoses or major procedures. This risk stratification by MDC is to ensure that the wide range of inpatient care and the different clinical factors that affect resource use are accounted for in the model. Each MDC corresponds to an organ system (e.g., MDC 2 covers diseases and disorders of the eye) or cause for admission (e.g., MDC 22 comprises MS-DRGs related to burns).

The measure also includes a Prior Inpatient Admission risk adjustor to ensure accurate cost comparison between episodes with and without prior inpatient admissions. as episodes where an inpatient stay occurs in the 30 days prior to the episode trigger are considered re-admissions that tend to be riskier and more resource-intensive than admissions.

Social Risk Factors: According to a 2014 National Quality Forum report,[35] the mechanisms underlying differences in resource use by socioeconomic status and race are complex and may be impacted by factors such as financial resources, community resources, historical and current discrimination, and reduced access to preventive services. Provider assumptions or implicit biases may impact quality of care for beneficiaries of different races. These factors may result in inefficient care, increased disease severity, or greater morbidity, leading to higher Medicare spending for beneficiaries depending on socioeconomic status or demographic status. Other social risk factors identified by the literature that can affect resource use include income, insurance (e.g., Medicaid), education, race and ethnicity, sex, social relationships, combinations of these factors, and residential and community context including rurality.[36,37,38]

---

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

[33] 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

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

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

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

[37] Assistant Secretary of Health and Human Services for Planning and Evaluation. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. Washington, D.C. December 2016

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

The MSPB Hospital measure was endorsed by NQF in 2013 and the measure methodology did not include social risk factor adjustment. The measure was reviewed again by NQF for re-endorsement in 2016/7 while under the NQF's trial period for risk adjustment of social risk factors and was endorsed without the addition of social risk factors. For these evaluations, we conducted analyses that demonstrated a small impact on measure scores from SRF inclusion and that the effect may be capturing patient level variation and provider level variation. Indeed, while acknowledging the "small effect" of SRF on the MSPB Hospital measure, the NQF "generally agree[d] the risk-adjustment method used in these measures met the NQF criteria given the data available to the developer and the measure testing results presented", "strongly urged the developer to continue testing additional variables within the risk-adjustment approach", and noted a preference for community-level SDS factors when individual factors are difficult to capture.[Error! Bookmark not defined.] Further, NQF's Risk Adjustment Expert Panel classified the MSPB Hospital measure as having a "Conceptual Relationship & Basis for Conceptual Relationship" with, and "Significant Association" to, social risk factors. [39]

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

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

Social risk factors were examined relative to the base model set of risk adjustment variables from the CMS-HCC V22 2016 model, disability status, ESRD status, interaction variables, and recent long-term care use, and in a step-wise fashion to determine the potential value of each social risk factor considered. Section 2b3.4b presents results on SRF testing.

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

- ☒ **Published literature**
- ☒ **Internal data analysis**
- ☐ **Other (please describe)**

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

The literature has extensively tested the use of the HCC model as applied to Medicare claims data. Although the variables in the HCC model were chosen to predict annual cost, CMS has also used this risk adjustment model in a number of other settings (e.g., ACOs, previous physician QRUR programs). Recalling that the risk

[39] NQF 2017 Evaluation of the NQF Trial Period for Risk Adjustment for Social Risk Factors

model relies on the existing CMS-HCC model, testing results for factors included in the CMS-HCC V22 2016 model can be found in the Pope et al (2011) report and the December 2018 CMS Report to Congress on risk adjustment in Medicare Advantage. [33,40]

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

Previous Response (2016): 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 (e.g., the transition from ICD-9 to ICD-10 codes) and is exhaustive on these code sets. Because the CMS-HCC model has already been extensively tested, we focus on adapting the CMS-HCC model to the MSPB-Hospital measure methodology.[33]

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. A larger proportion of low-income Medicare beneficiaries tended to use inpatient services in a given year compared to patients with higher incomes (25% and 17%, respectively). Lower-income beneficiaries are also twice as likely to use home health services as Medicare beneficiaries earning higher incomes.[41] End-of-life care for Medicare beneficiaries who are Black or Hispanic is substantially different than the end-of-life hospital services that Medicare beneficiaries who are White receive. Much of the variation in end-of-life care is due to differences in utilization levels among hospitalized patients. Beneficiaries who are Black and 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 convers, mechanical ventilation, and gastrostomy for artificial nutrition.[42]

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. These factors may result in inefficient care, increased disease severity, or greater morbidity,[43] leading to higher Medicare spending for beneficiaries depending on socioeconomic status or race.

---

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

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

[42] 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.

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

Given the conceptual and empirical relationship between income, race, and resource use, we analyzed both socioeconomic status (SES) and sociodemographic status (SDS), where SDS is defined as SES and race considered together. To determine SES, we used the United States Census Bureau's 2014 American Community Survey (ACS) 5-year estimates. The ACS dataset "Ratio of Income to Poverty Level of Families in the Past 12 Months" contains variables that provide population estimates of ranges of income-to-poverty ratios by ZIP code. Because individual family members may pool financial resources to provide care for older relatives, we used family income-to-poverty ratio in SES analysis instead of individual income-to-poverty ratio to better represent household decisions.[44] For a given ZIP code, the family income-to-poverty ratio dataset contains the variables: "Under .50", ".50 to .74", ".75 to .99", "1.00 to 1.24", "1.25 to 1.49", "1.50 to 1.74", "1.75 to 1.84", "1.85 to 1.99", "2.00 to 2.99", "3.00 to 3.99", "4.00 to 4.99", and "5.00 and over". Each of these variables gives the count of families in a given ZIP code whose income falls into that category range of income-to-poverty level. To illustrate, if the value for the ".50 to .74" variable is 10,000 for a particular ZIP code, that means that 10,000 families in that ZIP code have incomes that are between 50% and 74% of the federal poverty threshold.

The Enrollment Database (EDB) provided data on beneficiary race, and we look at race because race tracks with SES, and we wanted to see the impact on hospitals' performance on the MSPB-Hospital measure. While the EDB provides data on all race categories, there are concerns with the validity of the race categories other than Black and White (e.g., Asian, Hispanic, North American Native) due to underreporting in those categories.[45] As a result, we categorized beneficiaries as Black or Non-Black, where Non-Black is defined as all other race categories. The EDB also provided the ZIP codes for beneficiaries included in the sample. We then linked these beneficiary ZIP codes to the ACS ZIP code-level data on family income-to-poverty ratio to estimate the income-to-poverty ratio for each beneficiary with an MSPB-Hospital episode.

Using these data, we conducted a number of analyses related to disparities by population group. For race categories, we produced an estimated distribution of beneficiaries by income ratio (see Section 2b4.4b. for analysis). Additionally, we sought to determine the effect of incorporating SES or SDS into our risk adjustment model by determining the difference in MSPB-Hospital measure scores when including SES or SDS. We also analyzed correlation between MSPB-Hospital measure scores calculated with and without SES or SDS. The outcome of these analyses is discussed in Section 2b4.5.

Previous Response (2013): To account for case-mix variation and other factors, the MSPB risk-adjustment methodology broadly follows the CMS-HCC risk-adjustment methodology, which CMS uses to estimate Medicare Advantage (MA) premium adjustments.[46] Medicare also uses the HCC model to risk-adjust spending in: the Shared Savings Program Accountable Care Organizations (implemented in 2012) and the Medicare Physician Quality and Resource Use Reports (implemented in 2009). The accuracy of the ICD-9 codes used to

[44] Deaton, Angus S. and Paxson, Christina. Chapter 6: Measure Poverty among the Elderly. (Inquiries in the Economics of Aging, University of Chicago Press, January 1998), 171. https://core.ac.uk/download/pdf/6870973.pdf

[45] Zaslavsky, Alan M, John Z Ayanian, and Lawrence B Zaborski. "The Validity of Race and Ethnicity in Enrollment Data for Medicare Beneficiaries." Health Services Research 47.3 Pt 2 (2012): 1300–1321. PMC. Web. 28 Oct. 2016.

[46] Centers for Medicare and Medicaid Services, Office of the Actuary. "Announcement of Calendar Year (CY) 2009 Medicare Advantage Capitation Rates and Medicare Advantage and Part D Payment Policies." April 2008. http://www.cms.gov/MedicareAdvtgSpecRateStats/Downloads/Announcement2009.pdf

create HCCs has also been evaluated in previous studies, and all studies found high positive predictive values for Medicare claims-based diagnosis of acute myocardial infarction (AMI), chronic kidney disease (CKD), heart failure, coronary artery disease, diabetes, hypertension, and stroke with a diagnosis based on structured hospital record review.[47,48,49] A 2003 study found that CMS "administrative data was found to have diagnoses and conditions that were highly specific but that vary greatly by condition in terms of sensitivity."

Severity of illness is measured using 70 HCC indicators derived from the beneficiary's claims during the period 90 days prior to the start of the episode, an indicator of whether the beneficiary recently required long-term care, as well as the MS-DRG of the index hospitalization. The MSPB risk-adjustment methodology also includes status indicator variables for whether the beneficiary qualifies for Medicare through Disability or End-Stage Renal Disease (ESRD) and whether a beneficiary resides in a long-term care facility. Because the relationship between comorbidities' episode cost may be non-linear, the model includes interactions between HCCs and/or enrollment status variables. The MSPB risk-adjustment method does not control for the beneficiary's sex and race, but does include 12 age categorical variables. For a complete list of MSPB risk-adjustment variables, see the "MSPB Measure Information Form" available on QualityNet at the link provided in S.1.

All explanatory variables are calculated during the 90 days prior to the start of an episode. Calculating all health status variables prior to the start of an episode avoids the endogeneity problem which could occur if the diagnosis codes a hospital uses are included in the risk-adjustment model. Using claims data during the episode would incentivize hospitals to inflate the number of co-morbidities (i.e., number of diagnosis codes) that a beneficiary has to make their health status appear worse.

The MSPB risk-adjustment methodology (along with the entire MSPB methodology) was also put through official notice and comment rulemaking. The majority of commenters supported the risk adjustment for age and severity of illness. Some suggested further adjustment for race, sex, or socioeconomic factors, but Acumen and CMS opted to maintain consistency with the NQF's position against adjusting for these factors.

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

We analyzed the effect and impact of several social risk factors and the extent to which these effects may be attributable to hospitals relative to the patients they serve. We found inconsistency in the beneficiary-level estimates of the social risk factors, minimal impact to MSPB Hospital scores from SRF inclusion, and statistically significant hospital-level effects when decomposing the effects of select social risk factors.

---

[47] Kiyota, Uka, et al. "Accuracy of Medicare Claims-Based Diagnosis of Acute Myocardial Infarction: Estimating Positive Predictive Value on the Basis of Review of Hospital Records." American Heart Journal. 148(1): 99-104, July 2004.

[48] Winkelmayer, W. C., et al. "Identification of Individuals with CKD from Medicare Claims Data: A Validation Study." Am J Kidney Dis. 46(2): 225-232, Aug 2005.

[49] Birman-Deych, Elena, et al. "Accuracy of ICD-9-CM Codes for Identifying Cardiovascular and Stroke Risk Factors." Medical Care. 43(5): 480-485, May 2005.

We analyzed race, sex, dual status, income, education, unemployment, the AHRQ SES Index, and the Area Deprivation Index (ADI) as social risk factors. Beneficiary sex and dual status were obtained from the EDB and CME. Information on income, education, unemployment, and the AHRQ SES Index was obtained from ACS data. The ADI was constructed from 2015 ACS data by the University of Wisconsin School of Medicine and Public Health. Approximately 1.7 percent of beneficiary episodes with missing income and employment ACS data were excluded from the study[50] and the 15.92 percent of beneficiary episodes with missing ADI information were coded with a missing variable to observe any systematic effects of this population.

The percentage of female beneficiaries range from 27.0 percent to 63.5 percent across the 23 of the 26 MDCs in this measure that reasonably occur for both sexes (MDC 13 and MDC 14 are nearly 100 percent female as they are related to pregnancy, childbirth, and the female reproductive system, while MDC 12 is 0 percent female as it is related to the male reproductive system). For 23 out of 26 MDCs, most beneficiaries (55.7% - 84.4%) have non-dual status. The MDCs with a minority of non-dual status beneficiaries includes MDC 14 – Pregnancy, Childbirth, and the Puerperium (12.1%), MDC 25 – Human Immunodeficiency Virus Infections (30.1%), and MDC 19 – Mental Diseases and Disorders (44.0%). Income level is categorized into high, medium, and low from the continuous average income variable in ACS; therefore, each category has 33.3 percent of episodes. Approximately 2.0 to 8.1 percent of beneficiaries across all MDCs are classified as having below a high school education level, while 16.8 to 37.1 percent of beneficiaries have high unemployment designation (>10% for the Census Block Group).
Across all beneficiary episodes, the AHRQ Index ranged from 28.82 to 78.4 and approximately 14.36 of beneficiary episodes were ranked in the top quintile of the ADI's national ranking.

We examined the impact of including social risk factors into our risk adjustment model by running goodness of fit tests when different risk factors are added and compared to the base risk adjustment model, where the base risk adjustment model refers to the full standard set of risk adjustment variables from the CMS-HCC V22 2016 model, disability status, ESRD status, interaction variables, and recent long-term care use. We ran a step-wise regression to include the following additional social risk factors on top of the adapted base CMS-HCC model (Model 1). The models that added social risk factors to Model 1 in a step-wise manner include:

- Model 2: sex
- Model 3: dual status
- Model 4: sex + dual status
- Model 5: sex + dual status + race
- Model 6: sex + dual status + income + education + unemployment
- Model 7: sex + dual status + AHRQ SES Index
- Model 8: sex + dual status + race + income + education + unemployment
- Model 9: sex + dual status + race + AHRQ SES Index
- Model 10: sex + dual status + race + income + education + unemployment + AHRQ SES Index
- Model 11: sex + ADI Index Top Quintile
- Model 12: sex + dual status + ADI Index Top Quintile
- Model 13: sex + dual status + race + income + education + unemployment + ADI Index Top Quintile

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

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

First, we analyzed the model coefficients and p-values for each of the base and social risk factor models to understand whether any of the social risk factor covariates are predictive of episode cost. The model-specific T-tests and partial F-tests (relative to Model 1) indicated that social risk factors are likely predictive factors for determining resource use among beneficiaries for the relevant characteristic and MDC. For example, in models that include the AHRQ SES Index (models 7, 9, 10) and models that include the ADI (models 11, 12, 13), these indices have p-values less than or equal to 0.05 in at most 10 of the 26 MDC stratifications. Specifically, the AHRQ SES Index in Model 7 is statistically significant in 10 MDCs and in Model 11 is statistically significant in 6 MDCs.. The analysis also shows that the directions of the effects of social risk factors are not consistent. For example, low income episodes (as compared to high income episodes) and the AHRQ SES index may display both significant positive and negative coefficients of spending across MDCs. Considering the low-income categorization as an example, positive coefficients for low income may indicate that people with lower income tend to be more vulnerable and need additional resource use in their care. On the contrary, negative coefficients could indicate lower income people are expected to spend less, which may be a result of low-income patients having financial incentives to use less health care resources. They may be burden by co-pays for other services that they received covered by Medicaid. Appendix Tables 2b34b.a and 2b34b.b present these results.

Second, we analyzed the impact of adding social risk variables on overall model performance by looking at the differences in the O/E cost ratio with and without social factors in the risk adjustment model. When including social risk factors in our risk adjustment regression, minor differences in the O/E ratios, even for providers at high or low extremes of risk, indicates that social risk factor effects on the model performance are likely captured through existing risk adjustment variables. At least 99.5 percent of providers exhibited a ratio change less than ±0.03 across Models 1 – 13, with at least 93.2 percent of providers exhibiting a ratio change less than ±0.01. At least 84.2 percent of safety-net hospitals[51] and at least 91.7 percent of rural hospitals had ratio changes less than ±0.01, while at least 96.3 percent of non-safety net hospitals and at least 93.7 percent of urban hospitals showed comparable changes. All groupings exhibited a skewness to negative ratio changes, score improvements. Appendix Table 2b34b.c presents these results in detail.

We also analyzed the correlation between measure scores calculated with and without the social risk factors. The measure scores calculated with and without these social factors were highly correlated, ranging from 0.997 – 1.000 in Pearson and Spearman correlation coefficient. Appendix Table 2b34b.d presents these results in detail.

Models 1 through 13 suggested that the impact to measure scores from SRF inclusion was minimal and, perhaps more importantly, that effects of SRFs on predicting cost were ambiguous. Thus, we sought to clarify SRFs relationships by decomposing select social risk factors into their hospital- and beneficiary episode level effects. We decomposed these effects through random intercept models with contextualized beneficiary episode social risk factors and their hospital-level counterparts in the following models:

[51] Top quintile of DSH population nationwide

- Model A:
    - dual status (demeaned, episode level) + dual status proportions (hospital level)
- Model B:
    - AHRQ SES Index (demeaned, episode level) + AHRQ SES Index averages (hospital level)
- Model C:
    - ADI Index Top Quintile (demeaned, episode level) + ADI Index Top Quintile proportion (hospital level)

Of the 21 MDC models that had any statistically significant ($p<0.05$) decomposed dual factors, 19 MDCs had statistically significant hospital-level coefficients, with 15 of these hospital-level coefficients being statistically different from episode-level coefficients (Table 2b34b.e). Of the 16 MDC models that had any statistically significant ($p<0.05$) decomposed AHRQ Index factors, 12 MDCs had statistically significant hospital-level coefficients, with 10 of these hospital-level coefficients being statistically different from episode-level coefficients. Of the 15 MDC models that had any statistically significant ($p<0.05$) decomposed ADI Index factors, 14 MDCs had statistically significant hospital-level coefficients, with 12 of these hospital-level coefficients being statistically different from episode-level coefficients. These results indicate the presence of a provider-level effects (between-provider effects) and to the extent that these effects reflect true provider differences, including these social risk factors in risk-adjustment would mask these provider differences. Moreover, the inclusion patient/episode/community level SRFs alone may still partially adjust for provider-level differences.

Together, these results indicate that while social risk factors are a likely predictor of episode costs their inclusion would have a limited and inconsistent effect on measure scores and that some of the variation captured by tested covariates is attributable to provider-level differences.

Previous Response (2016):This section discusses the methodology used to analyze the following aspects of risk adjustment: (i) specification of the look-back period and stratification options, (ii) validity of current risk adjustment model, and (iii) evaluation of including SES and SDS.

Empirical evaluations of (i) focused on two specifications: first, the look-back period used to calculate comorbidities, and second, the methodology used to stratify the risk adjustment models. For the look-back period, the two options were 90-days, which is the period used in the current measure calculation, and 1 year. For stratifying the risk adjustment model, the options were to use only MDC, which is the current specification, or to use a combination of MDC and institutional status (i.e., whether a beneficiary is in long term care as determined using MDS data).

To demonstrate the validity of the MSPB risk adjustment methodology, we calculated the distribution of episode spending and R-squared 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 predictive ratio for each decile using the formula of average(expected cost)/average(observed cost) for all episodes in each decile. In addition, we calculated a "90/10 ratio," comparing the average cost of episodes in the first decile to the average cost of episodes in the tenth decile for observed costs and risk-adjusted costs. Risk-adjusted costs were calculated in two ways, by ratio and by residual. For the ratio calculation, we calculated risk-adjusted cost for each episode as (observed

cost/expected cost), multiplied by a national mean cost. For the residual calculation, we calculated risk-adjusted cost for each episode as (observed cost – expected cost) + national mean observed cost.

We examined the impact of including SES or SDS into our risk adjustment model with three tests: F-test of significance, difference in MSPB-Hospital measure scores, and correlation between MSPB-Hospital measure scores. First, we performed F-tests to assess the significance of SES and SDS on predicting resource use. The F-test revealed many significant p-values at the MDC level (see Appendix Table 2b4-4 and 2b4-6). This indicates that SES and SDS are likely predictive factors for determining resource use among beneficiaries for the relevant MDCs.

Overall, SES and SDS are likely predictive of variation in resource use. However, when including SES or SDS in our risk adjustment regression with other variables, the very minor change in hospital scores indicates that SES and SDS effects on hospital scores are largely captured through existing risk adjustment variables. We sought to determine the effect of incorporating SES or SDS into our risk adjustment model by determining the difference in MSPB-Hospital measure scores when including SES or SDS. In both cases, the differences in MSPB-Hospital measure scores were minimal (see Appendix Table 2b4-5 and 2b4-7). When including SES in risk adjustment, the MSPB-Hospital measure score for 97% of hospitals changed by ±0.01 or less. When including SDS in risk adjustment, the MSPB-Hospital measure score for 95% of hospitals changed by ±0.01 or less. Finally, we analyzed the correlation between MSPB-Hospital measure scores calculated with and without SES or SDS. The MSPB-Hospital measure scores calculated with and without SES were highly correlated (>0.998), as were measure scores calculated with and without SDS (>0.997). Because inclusion of SES and SDS factors has a minimal impact on the measure score and due to the high correlation values, we do not believe that including SES or SDS factors in the MSPB-Hospital risk adjustment methodology is appropriate.

Previous Response (2016) Appendix A: The MSPB-Hospital measure risk adjustment methodology is based on the CMS-HCC risk adjustment methodology, as described in the original measure submission. The measure uses OLS regressions for each Major Diagnostic Category to calculate expected episode cost.

The original submission included analysis on both socioeconomic status (SES) and sociodemographic status (SDS), where SDS is defined as SES and race considered together. The SES variable used was family income-to-poverty ratio, while race was calculated as non-black or black. Family income-to-poverty ratio was selected to strike a balance between individual and community factors related to SES, as individual family members may pool financial resources to provide care for older relatives. Empirical testing of SDS in the original submission included an F-test of significance, the difference in MSPB-Hospital measure score, and the correlation between measure scores calculated with SES/SDS and measure scores calculated without the SES or SDS variable.

NQF Committee Feedback: Committee members pointed out three areas that could require further investigation for risk adjustment. First, some members recognized that inclusion of SDS in risk adjustment has little impact on measure scores for the vast majority of providers, but expressed concern that not including SDS variables in risk adjustment could have a large impact on providers at the edge of the distribution. Second, the committee asked whether Acumen tested the use of a dual eligibility flag for risk adjustment to account for SDS factors. Third, one committee member noted that the disability variable taken from the enrollment file may not include the entire originally disabled population, as the Enrollment Data Base (EDB) contains a disability variable that turns from 1 to 0 when a beneficiary becomes 65 years old.

Methods: Acumen ran analyses using two versions of a risk adjustment model that account for dual status by including either (i) separate flags for full and partial dual status or (ii) one flag for full or partial dual status. The first analysis shows the distribution of the difference between a hospital's MSPB-Hospital measure score when calculating expected cost normally and when calculating expected cost using one of the two models accounting for dual status mentioned above. The difference was calculated as (MSPB-Hospital score) – (MSPB-Hospital score, adjusted for dual enrollment). Dual enrollment was identified using the dual enrollment variable in the Common Medicare Enrollment (CME) file.[52] Acumen also calculated the mean MSPB-Hospital measure score for providers based on the percent of beneficiaries with full or partial dual status. Finally, Acumen conducted an analysis of episodes for non-dual beneficiaries and episodes for dual beneficiaries to identify whether the measure score for hospitals that treat high proportions of dual beneficiaries are affected significantly by their dual population's episodes or not.

Acumen also investigated the originally disabled variable to identify whether beneficiaries older than the age of 65 had an active originally disabled flag (i.e., variable remained having a value of 1).

Results: Supplementary Table 5 shows that the standard deviation of the difference between provider's MSPB-Hospital measure score when calculated normally and when risk adjusting for dual beneficiaries is 0.003. The standard deviation does not change based on using separate flags for full and partial dual status or when using a single flag for full or partial dual status.

Supplementary Table 5: Difference in MSPB-Hospital Score When Risk Adjusting for Beneficiary Dual Status

Type of Dual Flag	# of Providers	Mean Difference	Std Dev.
Separate Flags for Full and Partial Dual	3,298	0.001	0.003
Single Flag for Full or Partial Dual	3,298	0.001	0.003

Supplementary Table 6 shows the distribution of the difference in MSPB-Hospital measure scores when adjusting for beneficiary dual status. When using separate flags for full dual status and partial dual status, the 0.1th percentile has a difference of -0.025, while the 99.9th percentile has a difference of 0.020. When using a single flag for full dual status or partial dual status, the 0.1th percentile has a difference of -0.029, while the 99.9th percentile has a difference of 0.017.

[52] Dual enrollment is defined using the dual_stus_cd variable in the CME file. Partial dual beneficiaries were defined as dual_stus_cd = 1, 3, 5, or 6, while full dual beneficiaries were defined as dual_stus_cd = 2, 4, or 7.

Supplementary Table 6: Distribution of Difference in MSPB-Hospital Score When Risk Adjusting for Beneficiary Dual Status

Type of Dual Flag	Percentiles: 0.1th	Percentiles: 1st	Percentiles: 5th	Percentiles: 25th	Percentiles: 50th	Percentiles: 75th	Percentiles: 95th	Percentiles: 99th	Percentiles: 99.9th
Separate Flags for Full and Partial Dual	-0.025	-0.005	-0.002	-0.001	0.000	0.002	0.005	0.010	0.020
Single Flag for Full or Partial Dual	-0.029	-0.005	-0.002	-0.001	0.000	0.002	0.005	0.010	0.017

Supplementary Table 7 shows the mean MSPB-Hospital measure score for hospitals, broken out by percent of beneficiaries with any episodes with full or partial dual status. There is not a clear trend for MSPB-Hospital measure scores based on the percent of dual-eligible beneficiaries, although hospitals with greater than 60% of beneficiaries having a dual-eligible episode have a higher MSPB-Hospital score on average. Supplementary Table 7 also shows the mean MSPB-Hospital measure score when restricting a hospital's episodes to their dual and non-dual episodes, respectively. These scores are similar, with scores for non-dual episodes being slightly higher for hospitals with 20 percent or higher of their beneficiaries having a dual episode.

Supplementary Table 7: MSPB-Hospital Measure Scores by % of Dual Beneficiary Episodes

Range	# of Providers	Mean MSPB-Hospital Measure Score	Std Dev.	Mean Score for Dual Episodes	Mean Score for Non Dual Episodes
1) 0% <= % of Beneficiaries with Any Full or Partial Dual Episode <= 10%	185	0.980	0.080	1.002	0.979
2) 10% < % of Beneficiaries with Any Full or Partial Dual Episode <= 20%	692	0.986	0.074	0.994	0.987
3) 20% < % of Beneficiaries with Any Full or Partial Dual Episode <= 30%	933	0.982	0.071	0.982	0.984
4) 30% < % of Beneficiaries with Any Full or Partial Dual Episode <= 40%	694	0.985	0.080	0.982	0.987
5) 40% < % of Beneficiaries with Any Full or Partial Dual Episode <= 50%	336	0.974	0.099	0.972	0.975
6) 50% < % of Beneficiaries with Any Full or Partial Dual Episode <= 60%	185	0.983	0.123	0.976	0.989

Range	# of Providers	Mean MSPB-Hospital Measure Score	Std Dev.	Mean Score for Dual Episodes	Mean Score for Non Dual Episodes
7) 60% < % of Beneficiaries with Any Full or Partial Dual Episode	186	1.021	0.168	1.018	1.037

Acumen's disability indicator is constructed from a field in the CMS EDB data, which obtains disability information using data from the CME file. Acumen reviewed the disability flag for beneficiaries who turned 65 between March 2015 and March 2016 and found that less than 0.001% of beneficiaries who had the disability flag in March 2015 did not have the disability flag in March 2016.

Interpretation: The first two analyses show that, in line with the original measure submission, there is a low impact of including the dual-eligibility flag as an SDS factor in risk adjustment. Most providers have a very minor change in their MSPB-Hospital measure score. In addition, the tails of the distributions are not disproportionately affected, as the overall magnitude of the change is low for almost all hospitals. The third analysis shows that specific hospitals are not affected by the inclusion of a dual enrollment flag, since measure scores do not vary much by percent of population that is dual status and measure scores are stable within a hospital across the dual and non-dual beneficiary populations. As such, including beneficiary dual status in the risk adjustment model has a minimal impact on MSPB-Hospital measure score. The recent ASPE report showed some differences in measure performance between hospitals with a high amount of Disproportionate Share Hospital payments and a low amount.[53] The analysis in Supplementary Table 7 suggests that these differences may be driven by hospitals with a very high concentration of dual eligible beneficiaries (above 60%), and that measure scores are high for both duals and non-duals in these hospitals. This suggests that these hospitals are relatively higher-cost hospitals for all types of patients.

In regard to the construction of the disability flag, the analysis shows that beneficiaries do not have their disability code reset when they turn 65. As such, the MSPB-Hospital measure's disability flag does continue to capture the originally disabled population. This implies that Acumen is using different information on disability from the problematic variable that the committee member identified.

Previous Response (2013): Because the CMS-HCC model has already been extensively tested, we focus on adapting the CMS-HCC model to the MSPB Measure methodology. To empirical evaluate the MSPB risk-adjustment methodology, we analyzed two specifications of the modified CMS-HCC risk-adjustment methodology by using R2 to measure model ability to explain variation: (1) evaluate the health status variables in the risk-adjustment by using one year of data prior to calculate comorbidities rather than 90 days; and (2) evaluate options for stratifying the risk-adjustment model (e.g., by MDC, MDC/Institutional Status). To demonstrate the validity of the MSPB risk-adjustment methodology, we (3) calculated the distribution of episode spending and R-squared by decile to examine the model's ability to predict both very low and high

[53] Office of the Assistant Secretary for Planning and Evaluation (ASPE). "Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs." December, 2016. Available at https://aspe.hhs.gov/system/files/pdf/253971/ASPESESRTCfull.pdf.

cost episodes. Specifically, we created a "risk score" for each episode calculated as the predicted values from each episode divided by the national average predicted value. After arranging episodes into deciles based on the risk score, we calculated the R-squared for each decile using the formula 1-(SSE/SST), where SSE = the sum of (episode observed spending – episode predicted spending) and SST = the sum of (episode observed spending – average overall observed spending).

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

In addition to other empirical validation (see Section 2b1.3), we conducted two analyses to assess the adequacy of our risk adjustment model.

First, we examined the variation in observed episode cost that is captured by our risk adjustment models and model parameter estimates. Specifically, we examined R-squared and adjusted R-squared fit statistics and model parameter estimates by Major Diagnostic Category (MDC). R-squared fit statistics summarize the extent to which the clinical factors included in the MSPB Hospital measures risk adjustment regression model explain variation in observed episode cost. These fit statistics should be interpreted with caution, as a low R-squared does not necessarily indicate that unexplained variation is attributable to variation in clinical care efficiency or vice versa. Further, individual risk adjustment coefficients and parameters should be viewed in the context of the entire model and set of MDCs, rather than being analyzed individually. For instance, coefficients indicate the incremental effect of a model variable, holding all other variables fixed. And, interactions between model variables must be interpreted in concert with the effects of those variables in isolation.

Second, we examined the model's ability to predict episode costs at varying levels of risk. As an episode's cost is expected to rise with a patient's clinical risk, we first allocated episodes into risk deciles by their ratio of expected cost to national average episode cost. Then, we examined the difference between a decile's average expected costs and average observed costs, the ratio between these costs (predictive ratio), and the average O/E cost ratio across deciles. A predictive ratio close to 1.0 is indicative of accurate cost prediction within a risk decile while an average O/E (or average E/O) ratio close to 1.0 is indicative of accurate cost prediction for the average episode within a decile.

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

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
The range of R-squared values for the MSPB Hospital measure risk adjust models, calculated by dividing explained sum of squares by total sum of squares, spanned from 0.11 to 0.67 across the MDCs. The adjusted R-squared range similarly spanned from 0.11 to 0.67.

Appendix Table 2b3.6.a provides the R-squared and adjusted R-squared values for each risk adjustment model. Appendix Table 2b3.6.b provides regression coefficients, standard errors and other statistics for each model.

Previous Response (2016): The average R-squared for the MSPB-Hospital measure risk adjustment model across all MDCs is 0.3014. The overall R-squared, calculated by comparing residuals to the difference between observed costs and the national mean cost across all MDCs, is 0.4757. Appendix Table 2b4-A also includes regression coefficients and standard errors for each of the covariates used in the risk adjustment models. More information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.[54]

Previous response (2013): The overall R-squared for the MSPB Measure risk adjustment model described in S.9.2. through S.9.4. is 0.4621. For your reference, the "Additional Information" Appendix beginning on page 24 of the "Scientific Acceptability" section also includes regression coefficients, standard error, and p-values of the covariates used in the risk-adjustment models. Recalling that the risk model relies on the existing CMS-HCC model, more information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.[55]

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
We interpret calibration as how accurately the risk model's predictions match the actual episode cost. We calculate the average O/E cost ratio for each risk decile to demonstrate the model's prediction accuracy for both high and low-cost episodes. The average expected cost differed from the average observed episode cost by 0.06 percent to 1.09 percent in absolute value across deciles. Further, both the predictive and O/E cost ratios were close to one, ranging from 0.99 to 1.01 across risk deciles.

Previous Response (2016):

1. *Evaluate options for look-back periods*: When changing the HCC "look-back" period from 90 days to 365 days: (i) 6.7% of episodes are dropped and (ii) the overall model fit (i.e., average of R-squared across all MDCs) decreases from 0.3014 to 0.2997. The R-squared, when calculated overall across MDCs, decreases from 0.4757 to 0.4736. More detailed statistics are shown in Appendix Table 2b4-1.
2. *Evaluate options for stratification of risk adjustment model*: When stratifying the risk adjustment model by MDC only, but with an indicator for institutional status (e.g., Long-Term Institutional (LTI) indicator) (current specification), the average R-squared across MDCs is 0.3014 and the overall R-squared is 0.4757. On the other hand, when stratifying the risk adjustment model by MDC, but with separate regressions for institutional and community beneficiaries, the average R-squared across MDCs is 0.3060 and the overall R-squared is 0.4778. In addition, when averaging across MDCs, 60.27% of regression variables have a p-value of less than 0.1 when using the MDC model, while only 48.93% of regression variables have a p-value of less than 0.1 when using the MDC/Institutional model. Further statistics by MDC are shown in Appendix Table 2b4-2.

Previous Response (2013):

1. Assessing the use of one year of data prior to the index admission to calculate comorbidities in the risk adjustment methodology rather than 90 days: When changing the HCC "look-back" period from 90 days to 365 days: (i) 6% of episodes are dropped (see Table 19 in the appendix) and (ii) the model fit

[54] Ibid.

[55] Ibid.

(i.e., R-squared) decreases from 0.4621 to 0.4601. The impact analysis also reveals that, despite the drop in episodes included and a decrease in model fit, most hospitals experience only a small change in their MSPB Measure values when switching the "look-back" period from 90 days to 365 days; in fact, Table 20 in the appendix shows that 78% of hospitals experience a gain or loss in the MSPB Measure values of less than 1 percentage point.

2. Evaluating options for stratifying the risk adjustment model (e.g., by MDC, MDC/Institutional Status): When stratifying the risk-adjustment model by MDC with a Long-Term Institutional (LTI) indicator (current specification), the R-squared is 0.4621. On the other hand, when stratifying the risk-adjustment model by MDC, but with separate regressions for institutional and community beneficiaries, the R-squared is 0.4645. When stratifying the risk-adjustment model by MDC, but with separate regressions for MDC type (i.e., MED, SURG), the R-squared is 0.4636. The MDC option was preferred because: (i) the improvement in R-squared is very small when moving to the MDC/Institutional Status specification and (ii) increasing the number of stratifications increases the risk of over-fitting, especially for MDCs with relatively few admissions.

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

Previous Response (2016): Evaluate the validity of the risk adjustment model: Table 1 below shows predictive ratios by risk decile for the MSPB-Hospital measure. The table shows that the model has consistent predictive ratios across risk score deciles, with the first decile having a predictive ratio of 0.994 and the tenth decile having a predictive ratio of 1.011.

Table 1: Predictive Ratios by Risk Decile for MSPB-Hospital

Decile	Number of Episodes	Average Observed Standardized Spending	Average Expected Standardized Spending	Predictive Ratio
1	542,061	$8,570.70	$8,621.74	0.994080
2	542,073	$11,166.26	$11,288.23	0.989192
3	542,060	$13,134.89	$13,136.85	0.999850
4	542,059	$15,066.59	$14,970.63	1.006413
5	542,063	$17,257.92	$17,122.40	1.007915
6	542,064	$19,242.71	$19,377.29	0.993055
7	542,064	$21,411.22	$21,642.11	0.989332
8	542,053	$24,151.26	$24,304.96	0.993676
9	542,072	$28,864.21	$28,920.72	0.998046
10	542,064	$46,105.10	$45,585.95	1.011388

The 90/10 ratio calculation shows that the risk adjustment model does effectively shrink the dispersion of the cost distribution. At the observed cost level, the 90/10 ratio is 6.22. The costs risk-adjusted by ratio have a 90/10 ratio of 3.40, and the costs risk-adjusted by residual have a 90/10 ratio of 3.21.

Previous Response (2013): Calculate the distribution of episode spending and R-squared by decile to show that the MSPB risk adjustment methodology does equally well predicting spending through all values of the model: The R-squared in the 3rd through 9th deciles are lower than overall R-squared in Table A below (includes outlier episodes) as well as Table B below (excludes outlier episodes). The R-squared in the 6th and 7th deciles are relatively low, ranging from approximately 1% to 3%. Additionally, the R-squared is always higher in Table B when outlier episodes are excluded.

Table A: Distribution of Spending and R-Squared by Decile* (Includes Outlier Episodes)

Decile	Episode Count	Min Risk Score	Max Risk Score	Avg. Obs Spending	Avg. Pred Spending**	Difference	R-Squared
1	446,268	-0.38	0.46	$7,442	$7,365	$77	0.7774
2	446,234	0.46	0.56	$9,607	$9,763	-$156	0.5861
3	446,197	0.56	0.65	$11,472	$11,506	-$34	0.3876
4	446,234	0.65	0.74	$13,379	$13,276	$103	0.2365
5	446,260	0.74	0.85	$15,164	$15,114	$50	0.1194
6	446,205	0.85	0.98	$17,452	$17,350	$101	0.0229
7	446,512	0.98	1.14	$20,047	$20,226	-$179	0.0100
8	445,951	1.14	1.31	$23,108	$23,237	-$128	0.0858
9	446,130	1.31	1.66	$27,830	$27,631	$199	0.1680
10	446,339	1.66	20.09	$45,115	$45,148	-$33	0.6903
TOTAL	4,462,330	-0.38	20.09	$19,062	$19,062	$0	0.4621

Note: *Decile are based on risk score calculated as ratio of predicted spending over national average predicted spending.

**Predicted spending is the predicted value from the regression.

Table B: Distribution of Spending and R-Squared by Decile* (Excludes Outlier Episodes)

Decile	Episode Count	Min Risk Score	Max Risk Score	Avg. Obs Spending	Avg. Pred Spending**	Difference	R-Squared
1	437,305	0.04	0.46	$7,087	$7,348	-$262	0.8644
2	437,313	0.46	0.56	$9,140	$9,730	-$590	0.6989
3	437,309	0.56	0.65	$10,905	$11,458	-$553	0.5135
4	437,248	0.65	0.74	$12,776	$13,213	-$436	0.3249
5	437,370	0.74	0.84	$14,596	$15,035	-$439	0.1744
6	437,310	0.84	0.98	$16,887	$17,247	-$360	0.0329
7	437,298	0.98	1.14	$19,566	$20,124	-$558	0.0140
8	437,320	1.14	1.31	$22,534	$23,144	-$609	0.1288
9	436,500	1.31	1.66	$27,237	$27,502	-$265	0.3627
10	438,118	1.66	20.17	$44,304	$45,039	-$735	0.7752
TOTAL	4,373,091	0.04	20.17	$18,506	$18,987	-$481	0.5978

Note: *Deciles are based on risk score calculated as ratio of predicted spending over national average predicted spending.

**Predicted spending is the Winsorized and renormalized predicted value.

2b3.9. Results of Risk Stratification Analysis:

N/A

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

The R-squared values, which measure the percentage of variation in results predicted by the model, are in line with or higher than the values presented in analyses of similar risk adjustment models.[33]

The average O/E cost ratios and the predictive ratios for all risk deciles are close to one. These results indicate that the model is accurately predicting spending, regardless of overall risk level. There was no evidence of excessive under- or over-estimation at the extremes of episode risk.

Previous Response (2016): The R-squared values for the model, which measure the percentage of variation in results predicted by the model, are in line with or are higher than the values presented in similar analyses of risk adjustment models.[56]

1. Evaluate options for look-back periods: As both the model fit and number of episodes included decrease when moving to a 365 day window for calculating comorbidities, the MSPB-Hospital risk adjustment model appropriately uses a 90 day period.
2. Evaluate options for stratification of risk adjustment model: These numbers justify the continued use of stratifying by MDC because: (i) the improvement in R-squared is very small when moving to the MDC/Institutional Status specification, (ii) increasing the number of stratifications by including

[56] Ibid, 6.

institutional status increases the risk of over-fitting, especially for MDCs with relatively few admissions, and (iii) more variables are statistically significant predictors in the MDC model as determined by a p-value of less than 0.1, which is generally accepted as statistically significant.

3. Evaluate the validity of the risk adjustment model: The risk decile table shows that the risk adjustment model has consistent predicted spending for all deciles. Predictive ratios close to 1 indicate that expected spending is accurately predicting observed spending. The maximum variation from 1 is in the tenth decile, with a predictive ratio of 1.011. Overall, this table shows that the model is accurately predicting observed spending, regardless of decile. A larger 90/10 ratio shows that the distribution of costs has a wider spread. This is an effective measure of dispersion, as compared to the standard deviation, because episode costs are skewed towards high-cost outliers. The 90/10 ratio, dropping by 45% and 48% for the ratio and residual calculations, respectively, does show that the risk adjustment for the MSPB-Hospital measure effectively reduces the dispersion in episode spending. Other investigations of the 90/10 ratio have found reductions of dispersion ranging from 20% to 48%.[57] This shows that the risk adjustment model does account for high-cost episodes and controls for the variance in observed spending.

Previous Response (2013):

1. Assessing the use of one year of data prior to the index admission to calculate comorbidities in the risk adjustment methodology rather than 90 days: When the FFS continuous enrollment requirement starts from 365 days prior to the start of the episode instead of 90 days prior to the start of the episode, there is no trade-off between the number of episodes included in the MSPB Measure and the model fit. In fact, both the number of episodes included and the model fit decrease (i.e., get worse).
2. Evaluating options for stratifying the risk adjustment model (e.g., by MDC, MDC/Institutional Status): The R-squared between the different options for stratifying the risk-adjustment model are comparable, indicating that the output is not very different. However, when separate regressions for the community/institutional model or the MED/SURG MDC model are run, degrees of freedom are lost and may cause over-fitting of the model.
3. Calculate the distribution of episode spending and R-squared by decile to show that the MSPB risk adjustment methodology does equally well predicting spending through all values of the model: Based on the distribution of spending and R-squared by decile, we believe that the MSPB risk-adjustment methodology is robust and fit consistently across deciles.

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

N/A

Previous Response (2016): N/A

[57] MaCurdy, Thomas et al. "Challenges in the Risk Adjustment of Episode Costs." CMS, February 2010. Available online at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Reports/downloads/MaCurdy_ERA_2010.pdf.

Previous Response (2013): Limited additional testing was performed because the MSPB Measure risk-adjustment methodology is intended to closely follow the established and extensively tested CMS-HCC risk-adjustment methodology. As previously discussed, however, we did test stratifying the model by MDC/Institutional Status rather than just stratifying the model by MDC. We also tested different look-back periods from the current 90 days.

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 examined the measure score distribution across all hospitals and by hospital characteristics. Specifically, we examined measure scores for all hospitals, hospital urban designation, hospital urban designation and bed size, Core-based statistical area (CBSA) region, ranges of disproportionate share hospital (DSH) patient percentages, safety-net designation - defined as the top quartile of DSH patient percentage distribution across hospitals nationwide, teaching hospital designation, and ranges of Medicare days as a percent of total inpatient days.

Given the distribution of Medicare payments across hospital characteristics in recent years and existing literature, we would expect hospitals with urban location designations to evidence higher MSPB Hospital measure scores relative to their counterparts.[62] We would also generally expect variation in hospital scores by hospital sizes – as measured by the number of inpatient beds – as this might be related to care provision inefficiencies,[58] and regional variation in service use.[59,60] Given the ambiguity surrounding teaching hospital costs,[61] the exclusion of IME add-on payments in the standardized Medicare payment that the MSPB Hospital measure uses, and that IME add-on payments largely account for teaching hospitals' larger aggregate margins (relative to non-teaching hospitals),[62] the relationship between MSPB Hospital measure scores and teaching hospital designation is unclear. Similarly, given that DSH patient proportion add-on payments are excluded in

---

[58] Giancotti, M., Guglielmo, A., Mauro M.. "Efficiency and optimal size of hospitals: Results of a systematic search", Plos One, March 29, 2017. Available online at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174533#abstract0

[59] Skinner, Jonathan et al. "A New Series of Medical Expenditure Measures by Hospital Referral Region: 2003-2008". The Dartmouth Institute for Health Policy and Clinical Practice. June 21, 2011. http://www.dartmouthatlas.org/downloads/reports/PA_Spending_Report_0611.pdf

[60] "Report to Congress: Regional Variation in Medicare Part A, Part B, and Part D Spending and Service Use", CMS, September, 2017. Available online at http://medpac.gov/docs/default-source/reports/sept17_regionalvariation_report_final_sec.pdf?sfvrsn=0.

[61] Burke LG, Khullar D, Zheng J, Frakt AB, Orav EJ, Jha AK. Comparison of Costs of Care for Medicare Patients Hospitalized in Teaching and Nonteaching Hospitals. JAMA Netw Open. 2019;2(6):e195229. doi:10.1001/jamanetworkopen.2019.5229

[62] Chart 6.13 and Chart 6.19 from "A Data Book: Health Care Spending and the Medicare Program", CMS, June 2019. Available online at: http://medpac.gov/docs/default-source/data-book/jun19_databook_entirereport_sec.pdf?sfvrsn=0

standardized payments and patient clinical risk adjustors, it is unclear what differences, if any, might be apparent between hospitals with high DSH proportions or Safety-net hospital designations and those without such patient populations.

Previous Response (2016): Our method to determine clinically meaningful differences in MSPB-Hospital measure scores consists of stratifying MSPB-Hospital measure scores by meaningful hospital characteristics, and comparing those results to expected findings discussed in the literature. Stratification is performed for each of the following characteristics: urban/rural location and hospital size; urban/rural location and geographic region;[63] and teaching status. We analyze the distribution of MSPB-Hospital measure scores for subgroups defined by these characteristics, as well as for the overall population. The purpose of this analysis is to ensure that MSPB-Hospital measure scores vary in a manner consistent with expectations. That is: the literature has identified certain characteristics with a meaningful relationship to hospital performance, and this analysis stratifies MSPB-Hospital measure scores by those same characteristics. This analysis is therefore slightly different than the reliability and validity analyses discussed in Sections 2a2 and 2b2, since it specifically seeks to confirm that the MSPB-Hospital measure behaves as expected with respect to well-documented and meaningful hospital characteristics.

Previous Response (2013): MSPB summary statistics include the percentile distribution of the MSPB score both overall and by hospital type (e.g., urban/rural status, bed size, region, teaching status). Although poor MSPB scores could be due to low quality care, it could also be the case that unobservable factors (e.g., large populations of patients for whom English is a second language, low adherence to treatment regimens) outside of hospitals' control make these hospitals perform worse. To identify hospitals that treat a large number of socioeconomically disadvantaged patients, the following analysis also classifies hospitals by their Disproportionate Share Hospital (DSH) percentage.[64]

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*)

Across all hospitals, the highest MSPB Hospital measure score is almost three and a half times the lowest MSPB Hospital measure score and the 90th percentile is just over twenty-one percent greater than the MSPB Hospital score at the 10th percentile (Table 2b4.2). The average MSPB Hospital measure score for hospitals in rural areas is approximately three percent lower than the average score for hospitals in large urban areas. The MSPB Hospital measure score varied across regions, with averages spanning 0.94 (West North Central) to 1.03 (West South Central). The average MSPB Hospital score for teaching hospitals is approximately 1.7 percent higher than non-teaching hospitals and safety-net hospital scores are approximately 1.4 percent higher than non-safety-net hospitals. Average scores by DSH percentage suggest an inverse parabolic relationship.

---

[63] The geographic regions used in this analysis are drawn from the census regions and divisions used by the U.S. Census Bureau. See "Census Regions and Divisions of the United States." U.S. Census Bureau. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

[64] The Medicare DSH patient percentage is equal to the sum of the percentage of Medicare inpatient days attributable to patients entitled to both Medicare Part A and Supplemental Security Income and the percentage of total inpatient days attributable to patients eligible for Medicaid but not eligible for Medicare Part A.

Previous Response (2016): Key findings include: (1) the highest single MSPB-Hospital measure score is more than five times higher than the hospital with the lowest MSPB-Hospital measure score; (2) the MSPB-Hospital measure score at the 90th percentile is almost 23 percent greater than the MSPB-Hospital score at the 10th percentile; (3) the average MSPB-Hospital measure score for rural hospitals is almost five percent lower than the average MSPB-Hospital measure score for urban hospitals; (4) the average MSPB-Hospital Measure score in the West South Central region is the highest for both urban and rural hospitals, followed by the Mid-Atlantic and New England for urban hospitals and the East South Central and East North Central for rural hospitals; and (5) the average MSPB-Hospital measure score for teaching hospitals is higher than the measure score for non-teaching hospitals. Appendix Tables 2b5-1 through 2b5-4 present these results.

Previous Response (2013): Key findings include: (1) the hospital with the highest MSPB score costs Medicare more than six times as much as the lowest cost hospital; (2) hospitals at the 90th percentile MSPB Measure cost Medicare 25 percent more per episode than hospitals at the 10th percentile; (3) rural hospitals out-perform urban hospitals; (4) the average MSPB Measure value in New England and the West South Central regions are the highest for both urban and rural hospitals; (5) teaching hospitals have higher average spending levels, but they also have higher expected spending amounts (due to a sicker patient case mix); and (6) hospitals with a large number of DSH-eligible patients are not significantly less efficient than hospitals with few DSH beneficiaries. Tables 15 through 18 in the appendix present these results.

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?
(i.*e., what do the results mean in terms of statistical and meaningful differences?*)
There is wide range and fairly symmetric dispersion of MSPB hospital scores that permit meaningful differentiation. Some score distributions, that one might expect to be comparable given standardized prices that exclude Medicare add-on payments, do indeed appear comparable. For example, non-teaching hospital average risk-adjusted episode cost is less than 1.7 percent lower than teaching hospitals. Moreover, while the extreme scores of safety-net hospitals (max: 1.68; min: 0.49) differ from those of non-safety-net hospitals (max: 1.59; min: 0.66), the score distributions between the 10th and 90th percentiles are quite similar. Additionally, score distributions across other hospital characteristics, vary in expected ways. For example, rural hospitals generally have lower measure scores relative to urban hospitals; and, regional variation is apparent.

Previous Response (2016): There exists clinically/practically significant variation in MSPB-Hospital measure scores, which indicates the measure's ability to capture differences in performance. There also exists significant variation in MSPB-Hospital measure scores when considered in light of certain clinically meaningful hospital characteristics. As noted above, rural hospitals tend to have lower MSPB-Hospital measure scores than urban hospitals, and the West South Central region has the highest average MSPB-Hospital measure score for both urban and rural hospitals. As mentioned in section S.11, low MSPB-Hospital measure score(s) indicates that the hospital or set of hospitals have low MSPB-Hospital amount(s) (i.e., risk-adjusted spending); measure scores less than 1 indicate that the MSPB-Hospital amount is less than the national episode-weighted median MSPB-Hospital amount across all hospitals during the given performance period. The results can be interpreted to mean that hospitals with lower MSPB-Hospital measure scores have lower risk-adjusted spending than other hospitals.

Our findings regarding variation in the MSPB-Hospital measure, particularly with respect to clinically meaningful hospital characteristics, are consistent with existing literature. Research by the Dartmouth Institute for Health Policy & Clinical Practice has found significant variation in hospital expenditures for the Medicare population,[65] which is consistent with our findings regarding significant variation across MSPB-Hospital measure score percentiles. Dartmouth has also found significant variation with respect to characteristics considered by our analysis. In particular, their research has found that southern and northeastern states generally have high Medicare utilization, and that certain urban areas had higher Medicare utilization.[66] These findings within the literature are consistent with our stratified findings of the MSPB-Hospital measure score by geographic region and urban/rural hospital. Other literature also found that academic centers tend to have higher Medicare spending, which is consistent with our findings about teaching hospitals.[67]

Previous Response (2013): There exists significant variation in spending relative to the typical hospital. For example, hospitals at the 90th percentile use 25 percent more resources per episode than hospitals at the 10th percentile. These figures also vary across hospital characteristics.

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

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

Note*: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

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

N/A

---

[65] Fisher, Elliott et al. "Health Care Spending, Quality, and Outcomes." The Dartmouth Institute for Health Policy and Clinical Practice. February 27, 2009. http://www.dartmouthatlas.org/downloads/reports/Spending_Brief_022709.pdf

[66] Skinner, Jonathan et al. "A New Series of Medical Expenditure Measures by Hospital Referral Region: 2003-2008". The Dartmouth Institute for Health Policy and Clinical Practice. June 21, 2011. http://www.dartmouthatlas.org/downloads/reports/PA_Spending_Report_0611.pdf

[67] Romley, John et al. "Spending and Mortality in US Acute Care Hospitals." Am J Manag Care. 2013;19(2):e46-e54

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A
2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
N/A

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
Since the MSPB Hospital measure is calculated using Medicare claims data, we expect a high degree of data completeness. To ensure further that we have complete and accurate data for each beneficiary who opens an episode, we exclude episodes where beneficiary date of birth information (an input to the risk adjustment model) cannot be found in the EDB or the beneficiary death date occurs before the episode trigger date.

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

2b6.2. what is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse;* ***if no empirical sensitivity analysis,*** *identify the approaches for handling missing data that were considered and pros and cons of each*)
The case count and select descriptive statistics for the four missing/errant data considered in the MSPB Hospital measure are shown in Table 8 (Appendix Table 2b2.2). These categories consistent of episodes where necessary beneficiary data, like the beneficiary age risk adjustor, was missing, or where competing or non-continuous insurance coverage made complete Medicare service use tallies potentially incomplete.

Table 8. Missing Data Categories for the MSPB Hospital Measure

Exclusion	# Episodes	Percent of Triggered Episodes	Average Observed Episode Cost	Average O/E Cost Ratio
Beneficiary date of birth is missing	9	0.00%	$46,013	0.58
Beneficiary death date occurred before the trigger date	64	0.00%	$15,975	0.67

Exclusion	# Episodes	Percent of Triggered Episodes	Average Observed Episode Cost	Average O/E Cost Ratio
Beneficiary has a primary payer other than Medicare during the episode window or in the 90-day lookback period	1,019,510	10.55%	$22,529	0.97
Beneficiary was not enrolled in Medicare Parts A and B, or was enrolled in Part C, during the 90-day lookback period and episode window	1,369,929	14.18%	$22,474	0.91

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 non-responders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted;* ***if no empirical analysis****, provide rationale for the selected approach for missing data*)

As the MSPB Hospital measure is calculated with Medicare claims data, we expect a high degree of data completeness, which is supported by the limited frequency of missing data for birth date and invalid beneficiary death date information above. Additionally, the measure removes beneficiaries that may have gaps in the Medicare claims history due to alternate enrollment. This data processing step ensures that we have complete and accurate information needed to calculate the measure.

Feasibility

F.1. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

F.1.1. Data Elements Generated as Byproduct of Care Processes.

Generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition

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

If other:

F.2. Electronic Sources

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

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

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

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

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

Attachment:

F.3. Data Collection Strategy

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

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

CMS uses Medicare administrative claims data that hospitals submit to CMS for payment to calculate the MSPB Hospital measure. As a result, the required data are readily available and retrievable without undue burden. These claims data used are maintained by CMS's OIS. These data undergo additional quality assurance checks during measure development and maintenance. Specifically, CMS has in place several hospital auditing programs used to assess overall claims code accuracy, ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analyses to identify potential problem areas and detect fraud. CMS also audits important data fields, including diagnosis and procedure codes, as well as other elements that are consequential to payment. Specifically, CMS works with Program Safeguard Contractors (PSCs)/Zone Program Integrity Contractors (ZIPCs) to ensure program integrity; the agency also uses Comprehensive Error Rate Testing (CERT) Contractors to ensure that Medicare payments are correct. Between 2005 and 2015, CERT estimates that proper payment, which is payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments each year, and 92.7 percent in FY2019.[1,2] CMS continues to perform successful corrective actions and give providers additional education to ensure accurate billing. To ensure claims completeness and inclusion of any corrections, the measure is calculated using data with a 3 month claims run-out from the end of the performance period. During the data preview for the MSPB Hospital measure, each hospital receives a Hospital-Specific Report (HSR) that provides information on the hospital's performance on the MSPB Hospital measure, as well as three supplementary hospital-specific data files (an index admission file, a beneficiary risk score file, and an MSPB Hospital episode file) related to the hospital's MSPB Hospital measure. Together, these files provide an overview of how the hospital performed on the MSPB Hospital measure as well as a summary of how hospitals in the state and in the nation performed. For example, each hospital's files provide the number of eligible admissions, average spending per episode, MSPB Hospital amount, and MSPB Hospital measure for the hospital as well as for the state and the nation. Additionally, each hospital's MSPB Hospital spending is broken into three categories (i.e., 3 days prior to index admission, during-index admission, and 30 days after hospital discharge), and within these categories, spending levels are broken down by claim type. For comparison, the state and national values for these breakdowns are given to hospitals as well. Further, each hospital's average observed spending and average expected spending (based on beneficiary age and health status) by Major Diagnostic Category (MDC) are presented in the hospital's HSR alongside analogous values at the state and national levels to allow the hospital to compare its case mix against the state and the nation. In addition to helping hospitals verify their MSPB Hospital measure scores and identify opportunities to improve efficiency, providing these files allows us to better communicate MSPB

Hospital scores to hospitals and allows hospitals to provide informed feedback to the measure contractor and CMS.

[1] Comprehensive Error Rate Testing (CERT) Program. "Appendices Medicare Fee-for-Service 2015 Improper Payments Report". Table A6. https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicare-FFS-Compliance-Programs/CERT/CERT-Reports-Items/Downloads/AppendicesMedicareFee-for-Service2015ImproperPaymentsReport.pdf

[2] Comprehensive Error Rate Testing Program. "2019 Medicare Fee-for-Service Supplemental Improper Payment Data" https://www.cms.gov/files/document/2019-medicare-fee-service-supplemental-improper-payment-data.pdf

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)?

There are no fees, licensing, or other requirements for use of the MSPB Hospital measure values and MSPB Hospital measure spending breakdowns made publicly available on Hospital Compare.

F.3.3. If there are any fees associated with the use of this measure as specified, attach the fee schedule here. (Save file as: F3_3_FeeSchedule)

Usability and Use

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

U.1.1. Current and Planned Use

Specific Plan for Use	Current Use (for current use provide URL)
Payment Program	Public Reporting

U.1.2. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Program Name: Hospital Value-Based Purchasing Program
Sponsor: CMS
Purpose: The Hospital VBP program provides financial incentives to subsection (d) hospitals based on their performance on selected quality measures. Section 1886(o)(2)(B)(ii) of the Social Security Act, 3001 of the Patient Protection and Affordable Care Act requires that CMS implement a measure of Medicare spending per beneficiary as part of it Hospital Value-Based Purchasing (VBP) initiatives. The hospital performance score for a performance period will be determined using a higher of its achievement or improvement score for the MSPB-Hospital measure as described in the FY 2012 IPPS Final Rule at 76 FR 51654-56. The MSPB Hospital measure score will be incorporated into the Hospital VBP Program as part of the Efficiency domain. Because the MSPB Hospital measure is the only measure currently in the Efficiency domain, the total points earned for the domain would be the points earned on the MSPB-Hospital measure. Each hospital´s Total Performance Score (TPS), used to calculate each hospital´s incentive payment, is calculated by combining its component domain scores. A hospital's improvement score is calculated from a comparison of the hospital's MSPB Hospital

measure value during a period of performance against the MSPB Hospital measure value during a baseline period.
Geographic Area and Number/Percentage of Patients: In the FY2020 Hospital VBP Program, 2,731 hospitals received adjustment factors that incorporated data MSPB Hospital measure data. The MSPB Hospital measure is reported publicly on CMS' Hospital Compare website. Number/Percentage of Patients: N/A

U.1.3. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A.

U.1.4. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

N/A.

U.2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation. How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Providers have a review and correction period during which they can view Hospital-Specific Reports (HSR) that contain information on the MSPB Hospital measure. The HSR provides information on the hospital's performance on the MSPB Hospital measure and cost breakouts of measure components in relation to state and national statistics. For example, each hospital's HSR provides the number of eligible admissions, average spending per episode, MSPB Hospital amount, and MSPB Hospital measure for the hospital as well as for the state and the nation. Additionally, each hospital's MSPB Hospital spending is broken into three categories (i.e., 3 days prior to index admission, during-index admission, and 30 days after hospital discharge), and within these categories, spending levels are broken down by claim type. For comparison, the state and national values for these category breakdowns are also provided in HSRs. Further, each hospital's average observed spending and average expected spending (based on beneficiary age and health status) by Major Diagnostic Category (MDC) are presented in the hospital's HSR alongside analogous values at the state and national levels to allow the hospital to compare its case mix against the state and the nation. HSR instructions and table footnotes are provided to assist providers with measure interpretation. Any hospital with an MSPB Hospital measure score that meets the 25-episode case minimum can request an HSR. The HSRs further include three supplementary hospital-specific data files (an index admission file, a beneficiary risk score file, and an MSPB Hospital episode file) that contain various data used to calculate episode costs, identify beneficiary risk factors, and Medicare claims used in measure calculation. Over the past two years, at least 79 percent of HSRs were downloaded by hospital providers. Section U.2.1.2 lists additional resources, including stakeholder outreach, for the MSPB Hospital measure.

We obtained feedback on the measure and potential refinements in February 2020 from a technical expert panel comprised 20 members with expertise in cost measure development and evaluation and quality improvement from diverse backgrounds, including clinicians, healthcare providers, academia, and patient advocacy organizations. The specific feedback and refinements adopted are discussed in Sections U.2.2.1-U.2.3.

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.

The HSRs are provided once per year, in early to mid summer, and providers may request re-uploads of their HSRs as needed. Further, CMS provides an annual webinar during which the MSPB Hospital measure methodology and measure score interpretation is detailed. The webinar includes a question & answer session and the transcript and recording of the webinar are posted publicly. CMS also provides email help-desk support for operations (HSR re-uploads) and other questions (e.g., methodological questions).

Further, the following materials are provided for educational/explanatory efforts on the public facing QualityNet website:

- Measure information form
- MSPB Hospital measure calculation example
- MSPB Hospital measure SAS documentation and code
- MSPB Hospital measure frequently asked questions document.

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.

Feedback, as obtained through annual question & answer sessions or email help-desk support, typically centers around methodological questions of clarification (e.g., which claims are included in the measure). Most of these questions come from providers, few questions come from researchers.

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

Potential refinements to the MSPB Hospital measure methodology that is in current use were identified from prior rule comments, past NQF endorsement cycles, and related measure development (e.g., MSPB Clinician). These potential refinements included

- Narrowing the Medicare costs and service use included in the measure
- Allowing readmissions to trigger new MSPB Hospital episodes
- Updating the MSPB Hospital measure's MSPB Amount (score numerator) calculation to evenly weight all of a hospital's episodes
- Additional social risk factors to consider for testing for social risk factor inclusion

These potential refinements were tested and reviewed by a Technical Expert Panel (TEP) in February 2020 as part of the MSPB Hospital measure's re-evaluation. The TEP comprised 20 members with expertise in cost measure development and evaluation and quality improvement from diverse backgrounds, including clinicians, healthcare providers, academia, and patient advocacy organizations.

Though no official vote was taken, panelists agreed that maintaining MSPB Hospital measure's holistic "all-cost" approach, allowing readmissions to trigger new MSPB Hospital episodes to increase measure surveillance, and updating the MSPB Hospital measure's MSPB Amount (score numerator) calculation to evenly weight all of a hospital's episodes were appropriate refinements. Panelists further provided additional considerations for ongoing social risk factor testing, like examining the impact of controlling for the Area Deprivation Index.

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

See U.2.2.2 – the technical expert panel noted included individuals from academia and advocacy organizations.

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

The ultimate status of the four potential refinements is as follows:

- Narrowing the Medicare costs and service use included in the measure

- Was not adopted. While such a narrowing of costs captured by the measure might result in improvements in the measure's statistical reliability, the added complexity from defining service/cost exclusions for the measure did not outweigh the potential improvement in reliability for an already highly reliable measure (see Section 2a2 of the Testing Attachment). Further, unlike the MSPB Clinician measure – a similar measure that does impose service/cost exclusions, for example, services surrounding the index admission were largely seen as in control of the hospital provider.

- Allowing readmissions to trigger new MSPB Hospital episodes
  - Was adopted. In addition to the benefit of increasing the measure's surveillance of beneficiary episodes, an indicator to identify readmission-based episodes in the risk adjustment model was included to ensure that episode costs for these types of episodes were accurately predicted.
- Updating the MSPB Hospital measure's MSPB Amount (score numerator) calculation to evenly weight all of a hospital's episodes
  - Was adopted. The overall impact on measure scores' was generally limited (e.g., less than 3 percent change in the overall score distribution end points), while allowing each risk-adjusted episode equal weight in a provider's measure score (see all hospital distribution in Section S.13.1 of this document)
- Additional social risk factors to consider for testing for social risk factor inclusion
  - Was not adopted. Testing of additional SRF factors, like the Area Deprivation Index, continued to exhibit minimal measure score impacts while suggesting the masking of provider-level effects, as in prior NQF cycle analyses (see Section 2b3 of the Testing Attachment).

The refinements that were adopted exhibited the intended impacts. For example, by allowing readmission inpatient stays to trigger new episodes in the MSPB Hospital measure, the number of episodes used in MSPB Hospital measure score calculations increased by 16.97 percent from 5.10 million to 5.97 million episodes (Appendix Table 3a). Further, the inclusion of an indicator variable to control for the readmission characteristic of an episode controlled for the higher observed cost of readmission-based episodes (mean: $26,552) relative to non-readmission episodes (mean: $21,565), as evidenced by average observed to expected episode cost ratios that are close to 1.00 and by differences between these average observed to expected episode cost ratios for readmission and non-readmission episode types that were largely less than 1 percent. Taken with the change in measure risk adjustment calculation that ensures equal weight of each risk-adjusted episode at a hospital, the MSPB Hospital measure refinements resulted in score changes of less than 3 percent, relative to the original measure methodology, for approximately 94.5 percent of providers (Appendix Table 3b).

The refinements adopted also further harmonized the MSPB Hospital measure with the MSPB Clinician measure. Section H2 provides more detail on this improved harmonization across MSPB measures.

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**

When comparing MSPB Hospital measure scores between 2017 and 2018, we see that nearly half of all hospitals improved on their MSPB Hospital measure score (more detail in Section IM.2.2.). The MSPB Hospital measure is able to effectively capture provider risk-adjusted spending during an episode and is able to capture differences between providers. Results from our testing are described in depth in the Testing Attachment included in this submission.

U.3.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the

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

N/A

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

No unintended consequences to individuals or populations have been identified during testing, and no evidence of unintended negative consequences to individuals or populations have been reported since implementation.

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

N/A

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)

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

Measure Name: Medicare Spending Per Beneficiary (MSPB) Clinician;

Measure Steward: CMS;

Measure Relationship to MSPB Hospital: The MSPB Hospital and MSPB Clinician measures are closely aligned. Both measures assess costs from the same time window (three days prior to the index admission to 30 days after discharge) and focus on the same target population of beneficiaries admitted to the inpatient setting. Together, these measures align the incentives for clinicians and hospitals taking care of Medicare patients who are hospitalized.

Measure Name: Medicare Spending Per Beneficiary (MSPB) PAC; Measure Steward: CMS; Measure Relationship to MSPB Hospital: 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.

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.2.1 Response: The MSPB Hospital measure has been harmonized with MSPB Clinician and MSPB-PAC in the following ways: (i) change in risk adjusted ratio calculation, and (ii) allowing readmissions to trigger an episode (specific to MSPB Clinician).

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.)

H.2.1 Response: The MSPB Hospital measure has been harmonized with MSPB Clinician and MSPB-PAC in the following ways: (i) change in risk adjusted ratio calculation, and (ii) allowing readmissions to trigger an episode (specific to MSPB Clinician).

The MSPB Hospital measure differs from MSPB Clinician and MSPB-PAC in that it captures all Medicare Part A and Part B costs associated with an episode that is triggered by an inpatient stay while MSPB Clinician, for example, excludes services that are unrelated to clinician care.

H.3.1 Response: The MSPB Hospital measure evaluates hospitals' efficiency relative to the efficiency of the median hospital. The target population is Medicare beneficiaries enrolled in Medicare Parts A and B who were discharged from short-term acute hospitals. There are currently no NQF-endorsed measures that address both this same measure focus and this same target population.

Contact Information

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

Co.2 Point of Contact: Ronique, Evans, kimberly.spaldingbush@cms.hhs.gov, 410-786-8882-

Co.3 Measure Developer if different from Measure Steward: Acumen, LLC

Co.4 Point of Contact: N/A, N/A, ccsq-macra-support@acumenllc.com

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

Technical Expert Panel Members:

Anita Bemis-Dougherty, American Physical Therapy Association

Kathleen Blake, American Medical Association

Akinluwa (Akin) Demehin, American Hospital Association

Kurtis Hoppe, American Academy of Physical Medicine and Rehabilitation

Caroll Koscheski, American College of Gastroenterology

Alan Lazaroff, American Geriatrics Society

Shirley Levenson, American Academy of Nurse Practitioners

Robert Leviton, American Medical Informatics Association

Edison Machado, American Health Quality Association

James Naessens, Mayo Clinic

Shelly Nash, Adventist Health System

Diane Padden, American Association of Nurse Practitioners

Parag Parekh, American Society of Cataract and Refractive

David Seidenwurm, American College of Radiology

Mary Fran Tracy, National Association of Clinical Nurse Specialists

Janice Tufte, Society for Participatory Medicine

Ugochukwu (Ugo) Uwaoma, Trinity Health of New England

Danny van Leeuwen, Health Hats

Michael Wasserman, California Association of Long Term Care Medicine

Adolph Yates, Jr., American Association of Hip and Knee Surgeons

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2012

Ad.3 Month and Year of most recent revision: 06, 2020

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? 06, 2021

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: Secondary CMS steward point of contact

Organization: Centers for Medicare & Medicaid Services

First Name: Helen

Last Name: Dollar-Maples

Email Address: Helen.Dollar-Maples@cms.hhs.gov

Phone Number: (410) 786-7214