Contents

Measures for Discussion (Brief)

Subgroup 1

- 3559 Hospital-Level, Risk-Standardized Improvement Rate in Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)
  [CMS/Yale/YNHH Center for Outcomes Research and Evaluation (CORE)]
  - Reliability: H-5; M-1; L-2; I-1 Passed
  - Validity: H-1; M-4; L-3; I-1 Consensus Not Reached
- 0715 Standardized adverse event ratio for children < 18 years of age undergoing cardiac catheterization (Boston Children's Hospital - Center of Excellence for Pediatric Quality Measurement)
  - Reliability: H-0; M-3; L-3; I-3 Not Passed
  - Validity: H-0; M-5; L-2; I-2 Consensus Not Reached
- 3556 National Healthcare Safety Network (NHSN) Nursing Home-onset Clostridioides difficile Infection (CDI) Outcome Measure (Centers for Disease Control and Prevention)
  - Reliability: H-0; M-1; L-5; I-3 Not Passed
  - Validity: H-0; M-1; L-5; I-3 Not Passed

- 3576 Pediatric Asthma Emergency Department Use (UCSF)
  - Reliability: H-0; M-0; L-7; I-2 Not Passed
  - Validity: H-0; M-2; L-4; I-3 Not Passed

Subgroup 2

- 2496 Standardized Readmission Ratio (SRR) for dialysis facilities (Centers for Medicare & Medicaid Services)
  - Reliability: H-0; M-4; L-3; I-0 Consensus Not Reached
  - Validity: H-0; M-2; L-5; I-0 Not Passed

Subgroup 3

- 3566 Standardized Ratio of Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge (ED30) for Dialysis Facilities (UM - Kidney Epidemiology and Cost Center)
  - Reliability: H-1; M-2; L-5; I-1 Not Passed
  - Validity: H-1; M-7; L-0; I-1 Passed
- 2539 Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy (Centers for Medicare & Medicaid Services)
  - Reliability: H-4; M-3; L-1; I-0 Passed
  - Validity: H-1; M-3; L-3; I-1 Consensus Not Reached

Measures that Passed (Not Pulled for Discussion) (Brief)

Subgroup 1

- 0076 Optimal Vascular Care (MN Community Measurement)
  - Reliability: H-5; M-3; L-1; I-0 Passed
  - Validity: H-3; M-3; L-2; I-1 Passed
  - Composite Construction: H-3; M-3; L-1; I-1 Passed
- 0716 Unexpected Complications in Term Newborns (California Maternal Quality Care Collaborative)
  - Reliability: H-5; M-3; L-0; I-1 Passed
  - Validity: H-3; M-4; L-1; I-1 Passed
- 2687 Hospital Visits after Hospital Outpatient Surgery (The Centers for Medicare & Medicaid Services)
  - Reliability: H-5; M-4; L-0; I-0 Passed
  - Validity: H-1; M-7; L-1; I-0 Passed

Subgroup 2

- 3561 Medicare Spending Per Beneficiary – Post Acute Care Measure for Inpatient Rehabilitation Facilities (Centers for Medicare & Medicaid Services)
  - Reliability: H-3; M-4; L-0; I-0 Passed

  - Validity: H-1; M-6; L-1; I-0 Passed
- 3562 Medicare Spending Per Beneficiary – Post Acute Care Measure for Long-Term Care Hospitals (Centers for Medicare & Medicaid Services)
  - Reliability: H-5; M-2; L-0; I-0 Passed
  - Validity: H-2; M-3; L-2; I-0 Passed
- 3563 Medicare Spending Per Beneficiary – Post Acute Care Measure for Skilled Nursing Facilities (Centers for Medicare & Medicaid Services)
  - Reliability: H-5; M-3; L-0; I-0 Passed
  - Validity: H-2; M-4; L-1; I-1 Passed
- 3564 Medicare Spending Per Beneficiary – Post Acute Care Measure for Home Health Agencies (Centers for Medicare & Medicaid Services)
  - Reliability: H-3; M-3; L-1; I-1 Passed
  - Validity: H-3; M-3; L-1; I-1 Passed
- 3574 Medicare Spending Per Beneficiary (MSPB) Clinician (The Centers for Medicare & Medicaid Services)
  - Reliability: H-1; M-4; L-3; I-0 Passed
  - Validity: H-0; M-5; L-3; I-0 Passed
- 3575 Total Per Capita Cost (TPCC) (The Centers for Medicare & Medicaid Services)
  - Reliability: H-1; M-6; L-0; I-0 Passed
  - Validity: H-1; M-4; L-2; I-0 Passed

Subgroup 3

- 0369 Standardized Mortality Ratio for Dialysis Facilities (Centers for Medicare & Medicaid Services)
  - Reliability: H-2; M-5; L-1; I-0 Passed
  - Validity: H-4; M-3; L-1; I-0 Passed
- 1463 Standardized Hospitalization Ratio for Dialysis Facilities (Centers for Medicare & Medicaid Services)
  - Reliability: H-2; M-6; L-1; I-0 Passed
  - Validity: H-3; M-5; L-1; I-0 Passed
- 2977 Hemodialysis Vascular Access: Standardized Fistula Rate (Centers for Medicare & Medicaid Services)
  - Reliability: H-4; M-5; L-0; I-0 Passed
  - Validity: H-1; M-7; L-1; I-0 Passed
- 2978 Hemodialysis Vascular Access: Long-term Catheter Rate (Centers for Medicare & Medicaid Services)
  - Reliability: H-4; M-5; L-0; I-0 Passed
  - Validity: H-1; M-6; L-2; I-0 Passed
- 3565 Standardized Emergency Department Encounter Ratio (SEDR) for Dialysis Facilities (UM - Kidney Epidemiology and Cost Center)
  - Reliability: H-2; M-6; L-1; I-0 Passed
  - Validity: H-1; M-5; L-3; I-0 Passed

Measures for Discussion (Detailed)

Measure #3559 Hospital-Level, Risk-Standardized Improvement Rate in Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA) (CMS/Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE)) (Consensus Not Reached)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** This patient-reported outcome-based performance measure will estimate a hospital-level, risk-standardized improvement rate (RSIR) following elective primary THA/TKA for Medicare fee-for-service (FFS) patients 65 years of age and older. Improvement will be calculated with patient-reported outcome data collected prior to and following the elective procedure. The preoperative data collection timeframe will be 90 to 0 days before surgery and the postoperative data collection timeframe will be 270 to 365 days following surgery.
- **Type of measure:** Outcome: PRO-PM
- **Data source:** Claims, Instrument-Based Data
- **Level of analysis:** Facility
- **Risk-adjustment:** Statistical risk adjustment with 19 risk factors
- **Sampling allowed:** No
- **Ratings for reliability:** H-5; M-1; L-2; I-1 → Measure passes
    - Reliability testing conducted at the data element and score level
    - Data element reliability testing assessed consistency and test-retest reliability of the Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR) and Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS, JR) instruments.
        - HOOS, JR internal consistency using Person Separation Index (PSI) was 0.86 and 0.87 in the two cohorts tested.
        - HOOS, JR test-retest results produced ICCs between 0.75 and 0.97.
        - KOOS, JR internal consistency using PSI was 0.84 and 0.85 in the two cohorts tested.
        - KOOS, JR test-retest results produced ICCs between 0.75 and 0.93.
    - Score level reliability testing consisted of a signal-to-noise analysis
        - Results from a sample of 123 hospitals yielded a mean of 0.95 and a range from 0.90 to 0.99.
    - Notes and results, concerns of SMP on reliability and specifications:
        - Measure specifications: Some NQF Panel members wanted clarification on the measure result calculation and definitions for “predicted,” “expected” and “overall observed” improvement.
        - Data element reliability: Concern was expressed about data element reliability testing for “critical data elements” other than the HOOS, JR and the KOOS, JR. Data elements of concern were noted to be those

that "make-up the denominator:" the two additional PRO tools used in the risk model and additional risk factors, including the clinical characteristics based on coding (e.g. liver disease, severe infection).
    - Reliability impact of proxy surveys: An NQF Panel member voiced concern about proxy assessment, noting that it "is unorthodox and can add significant noise."
    - Measure conversion: An NQF Panel member noted that the HOOS, JR and KOOS, JR appeared to have been transformed from 0-100 but no specifications on the approach to transformation were provided.
    - Score change calculation: An NQF Panel member noted that the interval over which the "change" in score appears to have been estimated (90-0 days prior to surgery and 270-365 days following surgery) is quite wide and could vary for an individual patient by as much as 6 months.
    - Exclusions: There was a request for clarification about how the measure accounts for patients that die between the hospital discharge and the postoperative PRO data collection period (270-365 days postoperatively), and whether they are considered "lost to follow-up." Another NQF Panel member noted that excluding deaths seemed reasonable but suggested a check on death as a possible adverse event.
- **Ratings for validity:** H-1; M-4; L-3; I-1 → Consensus not reached
    - Validity testing was conducted at both the data element and score levels
    - Data element validity testing included responsiveness, external validity, floor and ceiling effects for both HOOS, JR and KOOS, JR.
        - HOOS, JR responsiveness produced standardized response means relative to other PROMs (HOOS domains, The Western Ontario and McMaster University Arthritis Index [WOMAC] domains) measuring post-surgery hip improvement of 2.38 and 2.03 in the two samples.
        - HOOS, JR external validity used Spearman's correlation analysis with the HOOS and WOMAC instruments and produced 0.87 for both samples.
        - HOOS, JR showed floor (0.6%–1.9%) and ceiling (37%–46%) effects, and were comparable to or better than HOOS domains and the WOMAC.
        - KOOS, JR responsiveness produced standardized response means relative to other PROMs (KOOS, WOMAC) measuring post-surgery hip improvement of 1.79 and 1.70 in the two samples.
        - KOOS, JR external validity used Spearman's correlation analysis with the KOOS and WOMAC instruments and produced 0.89 and 0.91 for the two samples.
        - KOOS, JR showed floor (0.4%–1.2%) and ceiling (18.8%–21.8%) effects.
    - Score level validity testing included empirical comparisons to another quality measure: NQF 1550 Hospital-level risk-standardized complication rate (RSCR) following elective primary THA/TKA.
        - Comparison of THA/TKA PRO-PM RSIRs to RSCR categories indicates an increasing monotonic trend. Those hospitals in the "RSCR Worse than National Average" category have lower median RSIRs (51.87%) than the median RSIR (66.49%) of hospitals in the "RSCR Same as National Average" category, which is lower than that of hospitals in the "RSCR Better than National Average" category (71.13%).

- Notes and results, concerns of the SMP on Validity:
  - Attribution: An NQF Panel member noted concern about attributing changes in joint function to the hospital (versus care such as rehabilitation services) with a follow-up interval of nine months to one year following surgery.
  - "Unstaged procedures": An NQF Panel member suggested that the exclusion of staged procedures might eliminate up to 43% of procedures, and that the measure name should include "from unstaged procedures."
  - Exclusion analysis: An NQF Panel member noted concern that data were not provided on how the excluded patients impacted the performance measure scores.
  - Exclusion thresholds: Some NQF Panel members had questions about the 25-case volume threshold—what the threshold was based on, what happens to a facility that falls below the 25-case recommendation, if facilities without 25 cases would be excluded from the measure (and should be identified as an exclusion), and if excluded, whether it would create an incentive for them to not complete data.
  - Risk-adjustment: "The model was developed including cases from hospitals not used for reliability, validity, and missing data testing, i.e., hospitals with low caseloads (n<25) not recommended for this measure. Did the developers do a sensitivity test to assess the impact of excluding these hospitals from the risk-adjustment development sample on the risk-adjustment model?"
  - Meaningful differences: An NQF Panel member requested clarity for the data provided and whether there are meaningful differences between hospitals in the top quartile.
  - Missing data: Two NQF Panel members voiced concerns about missing data and that the only complete data were analyzed without accounting for what is likely "fairly extensive missingness." One of these members noted concern that missing surveys were accounted for but that missing responses within the survey were not.

Measure Developer Response

ITEMS TO BE DISCUSSED

- **Action items:**
  - Validity Testing
    - What are the most pressing concerns related to the validity of the measure?
    - The developer has provided responses to each of the concerns identified by the SMP. Would you like further clarification on the responses provided by the developer?
    - Revote on validity?

Measure #0715: Standardized adverse event ratio for congenital cardiac catheterization (Boston Children's Hospital - Center of Excellence for Pediatric Quality Measurement) (Not Passed)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: Jan 17, 2011 Most Recent Endorsement Date: Jun 29, 2015]
- **Description:** Ratio of observed to expected major adverse events (MAE) among patients undergoing congenital cardiac catheterization, risk-adjusted using the Catheterization for Congenital Heart Disease Adjustment for Risk Method II (CHARM II).
- **Type of measure:** Outcome
- **Data source:** Electronic Health Data, Electronic Health Records, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted;** Yes; No social factors included in the model
- **Sampling allowed**: No
- **Ratings for reliability:** H-0; M-3; L-3; I-3 → Measure does not pass
    - Reliability testing was conducted at the data element level by matching the facility records with the registry data. Testing was conducted across 13 centers (650 cases) and 96% accuracy was reported for 26/27 adverse events.
    - Reviewers noted concerns with the representativeness of the testing sample, lack of statistical testing beyond the percent accuracy noted above (96%), lack of testing for all critical data elements (including data elements in the denominator), and lack of score level testing
- **Ratings for validity:** H-0; M-5; L-2; I-2 → Consensus not reached
    - Developer reports empirical validity testing conducted at the measure score level. However, responses are focused on risk adjustment. Several reviewers noted concerns with this information and that it was inadequate for demonstrating score level validity.
    - One reviewer noted concerns with the inclusion of "procedure type" in the risk adjustment model and a lack of analysis of social risk factors.

Measure Developer Response

ITEMS TO BE DISCUSSED

- **Action items:**
    - Reliability Testing
        - Reliability must be demonstrated at each level of analysis specified in the measure.
        - Consider the additional information submitted by the developer:
            - Did the information provided by the developer on facility-level testing meet that requirement?
            - How were the other concerns raised by reviewers addressed?
        - Revote on reliability?
    - Validity Testing

- How do the results presented for validity testing demonstrate measure score validity? Are these results adequate to demonstrate the measure is sufficiently valid?
- Does the developer response adequately address the reviewers' concerns?
- Re-vote on validity?

Measure #3556 National Healthcare Safety Network (NHSN) Nursing Home-onset Clostirdioides difficile Infection (CDI) Outcome Measure (Not Pass)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** Standardized Infection Ratio (SIR) of nursing home facility onset incident Clostridioides difficile Infection (CDI) Laboratory-identified (LabID) events among all residents in the facility. Nursing home-onset incident CDI are defined as laboratory confirmed cases that develop four days after admission.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Records, Paper Medical Records
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes; No social risk factors included, includes 3 facility-level factors
- **Sampling allowed**: No sampling
- **Ratings for reliability:** H-0; M-1; L-5; I-3 → Measure does not pass
    - Data element validity testing was submitted in lieu of reliability testing. Per NQF guidelines this is acceptable and ratings for reliability should be based on ratings for data element validity.
    - There were concerns from multiple reviewers regarding the lack of reliability testing.
- **Ratings for validity:** H-0; M-1; L-5; I-3 → Measure does not pass
    - Validity testing was conducted at the data element level.
    - Sensitivity, Specificity, PPV and NPPV were calculated based on comparison of validators' and facilities' determination of the presence of a reportable CDI; testing results were based on 3 states (14 nursing homes)
        - Pooled Mean Sensitivity: 65.9%
        - Specificity: 68.5%
        - PPV: 84.3%
        - NPPV: 48.2%
    - Reviewers noted concerns with the variation in the validation method across states and its impact on the comparability of the results.
    - Reviewers also sought additional information on how the data element validity testing results related to each of the data elements (e.g., resident days, number of beds, patients admitted on CDI treatment), as this information was not included
    - Reviewers noted several concerns with the risk adjustment approach including lack of patient-level factors included in the model, concerns with adjusting away facility factors that may impact the quality of patient care, and inadequate evidence of risk model calibration and discrimination

ITEMS TO BE DISCUSSED

- **Action items:**
  - Reliability Testing
    - Given NQF's current guidance for data element validity testing, developers are not required to submit additional reliability testing.
      - Was the data element validity testing adequate to demonstrate validity?
  - Validity testing:
    - Would the panel provide some additional guidance to the developer on what would make the submission stronger?

Measure #2496 Standardized Readmission Ratio (SRR) for dialysis facilities (University of Michigan Kidney Epidemiology and Cost Center) (Not Pass)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: Dec 23, 2014 Most Recent Endorsement Date: Dec 09, 2016]
- **Description:** The Standardized Readmission Ratio (SRR) for a dialysis facility is the ratio of the number of index discharges from acute care hospitals to that facility that resulted in an unplanned readmission to an acute care hospital within 4-30 days of discharge to the expected number of readmissions given the discharging hospitals and the characterisstics of the patients and based on a national norm. Note that the measure is based on Medicare-covered dialysis patients.
- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes, no social factors included
- **Sampling allowed**: No
- **Ratings for reliability:** H-0; M-4; L-3; I-0 → Consensus not reached
  - Reliability testing conducted at the data element and measure score levels
    - Measure score reliability was demonstrated using IUR and PIUR, bootstrapping with resampling, to determine within and between facility variation and to identify outlier facilities
    - 2019 Results: IUR = 0.35 and PIUR = 0.61
    - Some reviewers noted concern that the IUR value has dropped significantly since last submission and there does not appear to be an explanation for why this occurred. The developer also did not provide the IUR in terms of sample size as it did with their prior submission.
    - Reviewers also noted concerns with the low values of the IUR and PIUR.
- **Ratings for validity:** H-0; M-2; L-5; I-0 → Measure does not pass
  - Validity testing conducted at the data element and measure score levels
    - The developer's (2009) description of data element validity testing does not meet NQF's requirements. Therefore, the Panel should focus their evaluation on the recent data submitted for measure score empirical validity testing.

- Measure score validity testing was demonstrated by comparing this measure to four other measures using Pearson's correlations (all statistically significant):
  - Standardized Hospitalization Ratio: r = 0.39
  - Standardized Mortality Ratio: r = 0.10
  - Vascular Access: Long-term catheter rate: r-0.04
  - Vascular Access: Standardized fistula rate: r = 0.06
- Reviewers expressed concern regarding the data presented for demonstrating meaningful differences; some state data is inadequate to fully evaluate this criterion.

Measure Developer Response

ITEMS TO BE DISCUSSED

- **Action items:**
  - Reliability testing
    - How should the IUR and PIUR results be interpreted? What is their relationship to one another, and should one be the basis for evaluation over the other? Is the IUR result sufficient to demonstrate reliability?
    - Does the developer's response for reliability adequately address the reviewers' for reliability concerns?
    - Revote on reliability?
  - Validity testing
    - How should the correlations for validity testing be interpreted?
    - Does the developer's response for validity adequately address the reviewers' for concerns?
    - Revote on validity?

Measure #3566 Standardized Ratio of Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge (ED30) for Dialysis Facilities (University of Michigan Kidney Epidemiology and Cost Center)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** The Standardized Ratio of Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge for Dialysis Facilities (ED30) is defined to be the ratio of observed over expected events. The numerator is the of the number of index discharges from acute care hospitals that are followed by an outpatient emergency department encounter within 4-30 days after discharge for eligible adult Medicare dialysis patients treated at a particular dialysis facility. The denominator is the expected number of index discharges followed by an ED encounter within 4-30 days given the discharging hospital's characteristics, characteristics of the dialysis facility's patients, and the national norm for dialysis facilities. Note that in this document, acute care hospital includes critical access hospitals and "emergency department encounter" always refers to an outpatient encounter that does not end in a hospital admission. This measure is calculated as a ratio but can also be expressed as a rate.

When used for public reporting, the measure calculation will be restricted to facilities with 11 eligible index discharges in the reporting year. This restriction is required to ensure patients cannot be identified due to small cell size

- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes, 87 factors, social factor (sex) is included
- **Sampling allowed**: None
- **Ratings for reliability:** H-1; M-2; L-5; I-1 → Measure does not pass
    - Reliability testing was conducted at the measure score level by calculating an inter-unit reliability (IUR) with bootstrapping; IUR = 0.451, PIUR = 0.570; Facilities with at least 11 eligible cases included.
    - Some reviewers expressed concerns regarding clarity of the specifications.
    - Reliability testing approach were found to be generally acceptable, but several reviewers noted the low/modest IUR results.
- **Ratings for validity:** H-1; M-7; L-0; I-1 → Measure passes
    - Validity testing was conducted by stratifying facilities into two categories: the 'better than/as expected' and 'worse than expected,' categories of SEDR. Calculated the mean score of several quality measures:
        - Standardized Mortality Ratio (SMR)
        - Standardized fistula Rate (SFR)
        - Standardized Transfusion Ratio (STrR)
        - Percentage of Prevalent Patients Waitlisted (PPPW)
        - Standardized Hospitalization Ratio (SHR)
        - Standardized ED Visit Ratio (SEDR)
    - Then compared mean performance scores across the two strata of 'better than/as expected' 'and 'worse than expected' categories.

Table 2. Classification of ED30 and mean facility performance scores for Related Measures, 2016-2017

	Facilities Missing	Better than/As Expected	Worse than expected	As Hypothesized?
SMR	412	1.00	1.05	Yes
STrR	717	0.99	1.21	Yes
SFR	510	63.32	63.64	No
PPPW	341	19.70	14.71	Yes
SRR	346	1.00	1.00	Yes
SEDR	205	0.99	1.49	Yes

- Risk model discrimination/calibration: c-statistic = 0.665; some reviewers noted concern with the decisions not to include social factors, but generally found the risk model to be adequate.

ITEMS TO BE DISCUSSED

- **Action items:**
    - Reliability Testing
        - Given the modest/low IUR result, are there other reliability tests that might help support the reliability of this measure?
        - What might the developer consider in retesting the measure or assessing its specifications to improve reliability?

Measure #2539 Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy (Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE)) (Consensus Not Reached)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: Dec 23, 2014 Most Recent Endorsement Date: Dec 23, 2014]
- **Description:** Facility-level risk-standardized rate of acute, unplanned hospital visits within 7 days of a colonoscopy procedure performed at a hospital outpatient department (HOPD) or ambulatory surgical center (ASC) among Medicare Fee-For-Service (FFS) patients aged 65 years and older. An unplanned hospital visit is defined as an emergency department (ED) visit, observation stay, or unplanned inpatient admission. The measure is calculated separately for ASCs, and HOPDs.
- **Type of measure:** Outcome
- **Data source:** Claims, Other
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes
- **Sampling allowed**: No
- **Ratings for reliability:** H-4; M-3; L-1; I-0 → Measure passes
    - Reliability testing conducted at the measure score level using a signal-to-noise analysis (Adams method); IQRs were also provided:
        - Median score for all HOPDs: 0.744 (≥30 procedures: 0.782)
        - Median score for all ASCs: 0.964 (≥30 procedures: 0.883)
    - These results were generally acceptable to reviewers and very few concerns were cited.
- **Ratings for validity:** H-1; M-3; L-3; I-1 → Consensus not reached
    - Validity was demonstrated at the measure score level by presenting face validity.
    - Based on NQF criteria, maintenance measures are required to submit empirical validity testing at the time of maintenance review, OR a rationale for why empirical validity testing could not be conducted.
    - The description of their systematic assessment of face validity using a TEP does meet NQF requirements. The Panel should consider whether the rationale provided by the developer for not conducting empirical analysis is sufficient.
        - 86% agreement among TEP members that the measure can be used to distinguish between providers.
        - Rationale provided cites difficulty finding an adequate comparator measure against which to compare for validity testing. The developer described a process by which they searched other NQF-endorsed measures focused on colonoscopy. “The three measures described [above] do not assess the domains of quality measured by the CMS colonoscopy measure. The facility-level scores for these measures would therefore not be expected to correlate with facilities’ 7-Day Risk-Standardized Hospital Visit rate and cannot be used to externally validate the CMS measure.”

- Reviewers also noted concerns with the lack of inclusion of social factors and the rationale for not including them. Two factors were considered: dual eligibility and AHRQ SES index, but ultimately neither was included in the model.
    - Developers analysis showed that ED patients and patients identified as low-SES using the AHRQ SES Index are at increased risk of post-colonoscopy hospital visits within seven days, even after adjusting for other risk factors in a multivariable model. "However, the scores estimated for facilities with and without either social risk factor are highly correlated. Importantly, there is no meaningful or systematic increase in measure scores for facilities with the highest proportion of patients with social risk factors. Further, the absolute increase in the risk of a hospital visit for patients with either of the two social risk factors is low…"

Measure Developer Response

ITEMS TO BE DISCUSSED

- **Action items:**
    - Validity Testing
        - Developers focused on 3 potential measures for comparators:
            - 1. Colorectal Cancer Screening (electronic clinical quality measure [eCQM]):
                - Identifies the proportion of patients in the recommended age group for colonoscopy screenings (50-75) who have had the procedure.
            - 2. Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients
                - Identifies the percentage of patients who have received a screening colonoscopy and have a regular recommended follow-up of ten years. This measure excludes patients who are older than 66 or who have a life expectancy of fewer than ten years, as the follow-up colonoscopy is no longer deemed beneficial. This measure is also not risk-adjusted.
            - 3. Colonoscopy Interval for Patients with a History of Adenomatous Polyps – Avoidance of Inappropriate Use
                - Measures the percent of patients who appropriately receive a colonoscopy more than three years after a previous colonoscopy. This measure is designed to track procedures that are inappropriately done within three years, and excludes procedures that occur within three years, but have a documented reason for the interval. This measure is not risk-adjusted.
        - Is the rationale for not submitting empirical analysis appropriate?
        - Are there other approaches or measures that should have been considered by the developer?
        - Does the developer response adequately address the reviewers' concerns?
        - Revote on validity?

Measure #3576 Pediatric Asthma Emergency Department Use (UCSF) (Not Pass)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** This measure estimates the rate of emergency department visits for children ages 3 – 21 who are being managed for identifiable asthma, using specified definitions. The measure is reported in visits per 100 child-years.
  The rate construction of the measure makes it a more actionable measure compared to a more traditional quality measure percentage construct (e.g., percentage of patients with at least one asthma-related ED visit). The rate construction means that a plan can improve on performance either through improvement efforts targeting all patients with asthma, or through efforts targeted at high-utilizers, since all visits are counted in the numerator. For a percentage measure, efforts to address high-utilizers will be less influential on performance and potentially have no effect at all even if a high utilizer goes from 8 visits a year to 1, since in order to improve performance, a high-utilizer has to get down to zero visits.
- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Health Plan
- **Risk-adjusted;** Yes; 6 factors, including 3 social factors (poverty level, education and unemployment)
- **Sampling allowed**: No
- **Ratings for reliability:** H-0; M-0; L-7; I-2 → Measure does not pass
  - Reliability testing conducted at the measure score level using ICC's (ranging between 0.00076-0.0029.
  - Reviewers noted some concerns with the complexity of the specifications in calculating age in child years as this could impact the reliable implementation of the measure.
  - Reviewers also noted concerns with the approach and results for testing reliability including-low ICC results which indicate greater with plan variation that between plans and that the ICC was not performed on randomly selected split samples.
- **Ratings for validity:** H-0; M-2; L-4; I-3 → Measure does not pass
  - Empirical validity testing was conducted at the measure score level by examining the impact of QI intervention using a difference-in-differences model using negative binomial regression. However, reviewers noted that this approach does not adequately demonstrate the validity of the measure, but rather that that the QI intervention has an impact on the outcome of interest.
  - Reviewers also noted that, although the measure is specified at the health plan level, the validity testing was done at the facility level.
  - Reviewers expressed concerns regarding the low R-squared value for the risk model indicating a very low predictive power, and it appeared that the model was validated on the same data set that was used to create it.

Measure Developer Response

ITEMS TO BE DISCUSSED

- **Action items:**
  - Reliability Testing
    - Does the developer's response adequately address the reviewers' concerns?
    - Revote on reliability?
  - Validity Testing
    - Does the developer's response adequately address the reviewers' concerns?
    - Revote on validity?

Appendix A: Measures that Passed (Not Pulled for Discussion) (Detailed)

Measure #0076: Optimal Vascular Care (MN Community Measurement)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure (Original Endorsement Date: Aug 10, 2009; Most Recent Endorsement Date: Dec 08, 2016)
- **Description:** The percentage of patients 18-75 years of age who had a diagnosis of ischemic vascular disease (IVD) and whose IVD was optimally managed during the measurement period as defined by achieving ALL of the following: Blood pressure less than 140/90 mmHg; On a statin medication, unless allowed contraindications or exceptions are present; Non-tobacco user; On daily aspirin or anti-platelet medication, unless allowed contraindications or exceptions are present**.**
- **Type of measure:** Composite
- **Data source:** Paper medical Records, electronic health record
- **Level of analysis:** Clinician Group
- **Risk-adjusted:** Yes; includes social factors (patient's insurance type and deprivation index)
- **Ratings for reliability:** H-5; M-3; L-1; I-0 → Measure passes with HIGH rating
  - Reliability testing conducted at the measure score level using signal to Noise analysis (Adams' method) = 0.809
- **Ratings for validity:** H-3; M-3; L-2; I-1 → Measure passes
  - Validity testing conducted at the score level by correlating the measure with other diabetes care measures.
    - While the reviewers questioned some of the assumptions made regarding the relationship between this measure and the comparators, they generally agreed the measure was valid.
    - Other reviewers raised concerns with a lack of clear validation results for the risk adjustment model.
- **Ratings for Composite:** H-3; M-3; L-1; I-1 → Measure passes

- Reviewers generally agreed the composite construct was valid, but did express concerns regarding the need for further analysis on the composite construct that would validate the composite on data collected since last endorsement.

Measure #0716: Unexpected Newborn Complications in Term Infants (California Maternal Quality Care Collaborative)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: Jan 17, 2011 Most Recent Endorsement Date: Oct 25, 2016]
- **Description:** This is a hospital level performance score reported as the percent of infants with Unexpected Newborn Complications among full term newborns with no preexisting conditions, typically calculated per year.
- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Facility, Integrated Delivery System, Population: Regional and State
- **Risk-adjusted:** No
- **Sampling allowed**: No sampling; but minimum case limit of 200 is recommended
- **Ratings for reliability:** H-5; M-3; L-0; I-1 → Measure passes
    - Reliability testing conducted at the measure score level using signal-to noise analysis (Adams' method)
        - 225 hospitals in California = 0.9 (200 case minimum)
    - Reviewers generally found this analysis to be acceptable.
- **Ratings for validity:** H-3; M-4; L-1; I-1 → Measure passes
    - Empirical validity testing conducted at the measure score and data element levels.
    - Developers used multiple approaches to demonstrate validity of the measure score and data elements:
        - Patient-level analysis to evaluate the associations between the Unexpected Newborn Complications measure and newborn length of stay and newborn hospital cost. Testing revealed statistically significant differences between length of stay and costs between patients with and without complications.
        - Hospital-level analysis to assess Pearson Correlation Coefficient between hospital rate of unexpected newborn complications and hospital average newborn length of stay (r = 0.4) and hospital average newborn cost (r = 0.37)
        - Comparison of the rate of unexpected newborn complications in the ICD-9 period (2014 and 2015 Jan to Sep) and in the ICD-10 period (2016 and 2017). No changes in rates were found.
        - Assessed Pearson Correlation Coefficient (r = 0.64) between hospital rate of unexpected newborn complications and hospital rate of NICU admission among these records.
    - Reviewers generally accepted these testing results as modest, but adequate to demonstrate validity

- Reviewers questioned the lack of risk adjustment, but found the measure generally acceptable given the developer's rationale: the extensive exclusions were intended to create a more homogenous denominator population. The developers also state they have accounted for social factors in the exclusions.

Measure #2687 Hospital Visits after Hospital Outpatient Surgery (CMS/Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE))

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: Sep 03, 2015 Most Recent Endorsement Date: Dec 08, 2015]
- **Description:** Facility-level, post-surgical risk-standardized hospital visit ratio (RSHVR) of the predicted to expected number of all-cause, unplanned hospital visits within 7 days of a same-day surgery at a hospital outpatient department (HOPD) among Medicare fee-for-service (FFS) patients aged 65 years and older.
- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes; 21 factors, no social factors included
- **Sampling allowed**: No
- **Ratings for reliability:** H-5; M-4; L-0; I-0 → Measure passes
    - Reliability testing conducted at the measure score level using signal-to-noise analysis (Adams' method) with IQRs; minimum 30 procedures: 0.839 (median); 0.756 (all facilities)
    - Reviewers noted few concerns with the clarity of the specifications (e.g., risk adjustment, qualifying events, exclusions), and generally agreed the testing approach was acceptable.
- **Ratings for validity:** H-1; M-7; L-1; I-0 → Measure passes
    - Validity testing conducted at the measure score level. The measure was compared with hospital-wide readmission rate (HWR) and results indicated a weak positive correlation as expected by developers (0.033, p = 0.07).
    - Developer also presented face validity results, however, because this is a maintenance measure, empirical validity testing should be the basis for evaluation.
    - Risk model discrimination and calibration: c statistic = 0.684; developer reports good discrimination and predictive ability based on risk decile plot.
    - Reviewers expressed concern, but generally accepted the validity testing results as a weak, but acceptable, demonstration of validity.

Measure Developer Response

Measure #3561 Medicare Spending Per Beneficiary – Post Acute Care Measure for Inpatient Rehabilitation Facilities (Centers for Medicare & Medicaid Services)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** The Medicare Spending Per Beneficiary – Post Acute Care Measure for Inpatient Rehabilitation Facility (MSPB-PAC IRF) was developed to address the resource use domain of the Improving Medicare Post-Acute Care Transformation Act of 2014 (IMPACT Act). This resource use measure is intended to evaluate each IRF's efficiency relative to that of the national median IRF. Specifically, the measure assesses Medicare spending by the IRF and other healthcare providers during an MSPB episode. The measure reports the ratio of the payment-standardized, risk-adjusted MSPB-PAC Amount for each IRF divided by the episode-weighted median MSPB-PAC Amount across all IRFs. The MSPB-PAC Amount is the ratio of the observed episode spending to the expected episode spending, multiplied by the national average episode spending for all IRFs. The measure is calculated using two consecutive years of Medicare Fee-for-Service (FFS) claims data and was developed using calendar year (CY) 2015-2016 data. This submission is based on fiscal year (FY) 2016-2017 data; i.e., IRF admissions from October 1, 2015 through September 30, 2017.
  Claims-based MSPB-PAC measures were developed in parallel for the IRF, long-term care hospital (LTCH), skilled nursing facility (SNF), and home health agency (HHA) settings to meet the mandate of the IMPACT Act. To align with the goals of standardized assessment across all settings in PAC, these measures were conceptualized uniformly across the four settings in terms of the construction logic, the approach to risk adjustment, and measure calculation. Clinically meaningful case-mix considerations were evaluated at the level of each setting. For example, clinicians with IRF experience evaluated IRF claims and then gave direction on how to adjust for specific patient and case-mix characteristics.
  The MSPB-PAC IRF measure was adopted by the Centers for Medicare & Medicaid Services (CMS) for the IRF Quality Reporting Program (QRP) and finalized in the FY 2017 IRF Prospective Payment System (PPS) Final Rule.[1] Public reporting for the measure began in Fall 2018 through the IRF Compare website (https://www.medicare.gov/inpatientrehabilitationfacilitycompare/) using FY 2016-2017 data.
  Notes: [1] Medicare Program; Inpatient Rehabilitation Facility Prospective Payment System for Federal Fiscal Year 2017 Federal Register, Vol. 81, No. 151. https://www.gpo.gov/fdsys/pkg/FR-2016-08-05/pdf/2016-18196.pdf;
- **Type of measure:** Cost/Resource Use
- **Data source:** Assessment Data, Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes, 146 factors; no social factors included
- **Sampling allowed**:
- **Ratings for reliability:** H-3; M-4; L-0; I-0 → Measure passes
    - Reliability testing was conducted at the measure score level using split-sample method with correlations (ICCs): Mean = 0.87; and signal-to-noise analysis (Adams' method): 0.86 (mean score); minimum 20 episodes in 2-year measurement period
    - Reviewers expressed some concerns with the specifications including lack of clarity, which may prevent others from reliably reproducing the measure. Concerns were regarding some of the steps to calculate the measure including exclusions, how

overlapping episodes are addressed, how events are ended, and how outlier exclusions are handled.
    - Reliability testing results were generally acceptable to reviewers.
- **Ratings for validity:** H-1; M-5; L-1; I-0 → Measure passes
    - Validity testing was conducted at the score level using multiple approaches:
        - Examined the correlation with known indicators of resource or service utilization: Hospital admissions and emergency room (ER) visits during the episode period. Compared the ratio of observed over expected spending for MSPB-PAC IRF episodes with and without hospital admissions occurring in the episode period; and compared the observed over expected spending for episodes with and without ER visits
        - Examined the correlation between MSPB-PAC IRF scores and the Discharge to Community (DTC) rates for FY 2016-2017; Pearson Correlation = -0.193
        - Examined the correlation between MSPB-PAC IRF scores and provider's scores on the Percent of Residents or Patients with Pressure Ulcers That Are New or Worsened (Short-Stay) measures (NQF #0678); Pearson Correlation = 0.1207
    - Risk model discrimination: adjusted = R-squared = 0.1157
        - Reviewers expressed concerns regarding the decision not to include social factors despite the conceptual and empirical analyses and the low R-squared value
    - One panel member expressed concern regarding the impact of poor risk adjustment on the validity and usability of the measure.

Measure Developer Response

Measure #3562 Medicare Spending Per Beneficiary – Post Acute Care Measure for Long-Term Care Hospitals (Centers for Medicare & Medicaid Services)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** The Medicare Spending Per Beneficiary – Post Acute Care Measure for Long-Term Care Hospitals (MSPB-PAC LTCH) was developed to address the resource use domain of the Improving Medicare Post-Acute Care Transformation Act of 2014 (IMPACT Act). This resource use measure is intended to evaluate each LTCH's efficiency relative to that of the national median LTCH. Specifically, the measure assesses Medicare spending by the LTCH and other healthcare providers during an MSPB episode. The measure reports the ratio of the payment-standardized, risk-adjusted MSPB-PAC Amount for each LTCH divided by the episode-weighted median MSPB-PAC Amount across all LTCH facilities. The MSPB-PAC Amount is the ratio of the observed episode spending to the expected episode spending, multiplied by the national average episode spending for all LTCHs. The measure is calculated using two consecutive years of Medicare Fee-for-Service (FFS) claims data and was developed using calendar year (CY) 2015-2016 data. This submission is based on fiscal year (FY) 2016-2017 data; i.e., LTCH admissions from October 1, 2015 through September 30, 2017.

Claims-based MSPB-PAC measures were developed in parallel for the LTCH, inpatient rehabilitation facility (IRF), skilled nursing facility (SNF), and home health agency (HHA) settings to meet the mandate of the IMPACT Act. To align with the goals of standardized assessment across all settings in PAC, these measures were conceptualized uniformly across the four settings in terms of the construction logic, the approach to risk adjustment, and measure calculation. Clinically meaningful case-mix considerations were evaluated at the level of each setting. For example, clinicians with LTCH expertise evaluated LTCH claims and then gave direction on how to adjust for specific patient and case-mix characteristics.
The MSPB-PAC LTCH measure was adopted by the Centers for Medicare & Medicaid Services (CMS) for the LTCH Quality Reporting Program (QRP) and finalized in the FY 2017 LTCH Prospective Payment System (PPS) Final Rule.[1] The measure entered into use on October 1, 2016. Public reporting for the measure began in Fall 2018 through the LTCH Compare website (https://www.medicare.gov/longtermcarehospitalcompare/) using FY 2016-2017 data.
Notes: [1] Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Policy Changes and Fiscal Year 2017 Rates. Federal Register, Vol. 81, No. 162. https://www.govinfo.gov/content/pkg/FR-2016-08-22/pdf/2016-18476.pdf;

- **Type of measure:** Cost/Resource Use
- **Setting of Care:** Post-Acute
- **Data source:** Assessment Data, Claims, Enrollment Data, Other Care
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes, 232 factors; no social factors included
- **Sampling allowed**:
- **Ratings for reliability:** H-5; M-2; L-0; I-0 → Measure passes
    - Reliability testing conducted at the measure score level using the signal-to-noise analysis (Adams method) (Mean score = 0.87) and split-sample testing approach using ICCs (Mean score = 0.86); minimum of 20 episodes in a 2-year measurement period
    - Reviewers expressed some concerns with the specifications including lack of clarity which may prevent others from reliably reproducing the measure. Their concerns were regarding some of the steps to calculate the measure including exclusions, how overlapping episodes are addressed, how events are ended, and how outlier exclusions are handled
    - Reliability testing results were generally acceptable to reviewers.
- **Ratings for validity:** H-2; M-3; L-2; I-0 → Measure passes
    - Conducted empirical validity testing at the score level using multiple approaches:
        - Examined the correlation with known indicators of resource or service utilization: Hospital admissions and emergency room (ER) visits during the episode period; compared the ratio of observed over expected spending for MSPB-PAC LTC episodes with and without hospital admissions occurring in the episode period; and compared the observed over expected spending for episodes with and without ER visits.
            - Mean observed to expected ratio without ER visit = 1.00
            - Mean observed to expected ratio with at least 1 ER visit = 1.02
        - Examined the correlation between MSPB-PAC LTCH scores and the Discharge to Community (DTC) rates for FY 2016-2017; Pearson Correlation = -0.2063

        - Examined the correlation between MSPB-PAC LTCH scores with four other quality measures; all correlations were weak and not statistically significant:
            - Percent of Residents or Patients with Pressure Ulcers That Are New or Worsened (Short-Stay) (#0678): r = -0.0927
            - Catheter-associated Urinary Tract Infection (CAUTI) Outcome Measure (#0138): r = 0.0435
            - Central line-associated Bloodstream Infection (CLABSI) Outcome Measure (#0139): r = 0.0074
            - Facility-wide Inpatient Hospital-onset Clostridium difficile Infection (CDI) Outcome Measure (#1717): r = -0.0335
        - Risk model calibration/discrimination: Overall adjusted R-squared = .04894
    - Reviewers expressed concerns regarding the decision not to include social factors despite the conceptual lack of clarity around the "site-neutral" prediction model and outlier exclusions.

Measure Developer Response

Measure #3563 Medicare Spending Per Beneficiary – Post Acute Care Measure for Skilled Nursing Facilities (Centers for Medicare & Medicaid Services)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** The Medicare Spending Per Beneficiary – Post Acute Care Measure for Skilled Nursing Facilities (MSPB-PAC SNF) was developed to address the resource use domain of the Improving Medicare Post-Acute Care Transformation Act of 2014 (IMPACT Act). This resource use measure is intended to evaluate each SNF's efficiency relative to that of the national median SNF. Specifically, the measure assesses Medicare spending by the SNF and other healthcare providers during an MSPB episode. The measure reports the ratio of the payment-standardized, risk-adjusted MSPB-PAC Amount for each SNF divided by the episode-weighted median MSPB-PAC Amount across all SNFs. The MSPB-PAC Amount is the ratio of the observed episode spending to the expected episode spending, multiplied by the national average episode spending for all SNFs. The measure is calculated using two consecutive years of Medicare Fee-for-Service (FFS) claims data and was developed using calendar year (CY) 2015-2016 data. This submission is based on fiscal year (FY) 2016-2017 data; i.e., SNF admissions from October 1, 2015 through September 30, 2017.
  Claims-based MSPB-PAC measures were developed in parallel for the SNF, long-term care hospital (LTCH), inpatient rehabilitation facility (IRF), and home health agency (HHA) settings to meet the mandate of the IMPACT Act. To align with the goals of standardized assessment across all settings in PAC, these measures were conceptualized uniformly across the four settings in terms of the construction logic, the approach to risk adjustment, and measure calculation. Clinically meaningful case-mix considerations were evaluated at the level of each setting. For example, clinicians with SNF experience evaluated SNF claims and then gave direction on how to adjust for specific patient and case-mix characteristics.

The MSPB-PAC SNF measure was adopted by the Centers for Medicare & Medicaid Services (CMS) for the SNF Quality Reporting Program (QRP) and finalized in the FY 2017 SNF Prospective Payment System (PPS) Final Rule.[1] Public reporting for the measure began in Fall 2018 through the Nursing Home Compare website (https://www.medicare.gov/nursinghomecompare/search.html using FY 2017 data.
Notes:
[1] Medicare Program; Prospective Payment System and Consolidated Billing for Skilled Nursing Facilities for FY 2017, SNF Value-Based Purchasing Program, SNF Quality Reporting Program, and SNF Payment Models Research; Final Rule. Federal Register, Vol. 81, No. 151. https://www.govinfo.gov/content/pkg/FR-2016-08-05/pdf/2016-18113.pdf

- **Type of measure:** Cost/Resource Use
- **Data source:** Assessment Data, Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes, 124 factors; no social factors included
- **Sampling allowed**:
- **Ratings for reliability:** H-5; M-3; L-0; I-0 → Measure passes
    - Reliability testing conducted at the measure score level using the signal-to-noise analysis (Adams method) (Mean score = 0.92) and split-sample testing approach using ICCs (Mean score = 0.93); minimum of 20 episodes in a 2-year measurement period
    - Reviewers expressed some concerns with the specifications including lack of clarity which may prevent others from reliably reproducing the measure. Their concerns were regarding some of the steps to calculate the measure including exclusions, how overlapping episodes are addressed, how events are ended, and how outlier exclusions are handled
    - Reliability testing results were generally acceptable to reviewers.
- **Ratings for validity:** H-2; M-4; L-1; I-1 → Measure passes
    - Conducted empirical validity testing at the score level using multiple approaches:
        - Examined the correlation with known indicators of resource or service utilization: Hospital admissions and emergency room (ER) visits during the episode period. Compared the ratio of observed over expected spending for MSPB-PAC SNF episodes with and without hospital admissions occurring in the episode period; and compared the observed over expected spending for episodes with and without ER visits.
            - Mean observed to expected ratio without ER visit=0.95
            - Mean observed to expected ratio with at least 1 ER visit =1.21
        - Examined the correlation between MSPB-PAC SNF scores and the Discharge to Community (DTC) rates for FY 2016-2017; Pearson Correlation = -0.3777
        - Examined the correlation between MSPB-PAC LTCH scores with quality measure: Percent of Residents or Patients with Pressure Ulcers That Are New or Worsened (Short-Stay) (NQF #0678): Pearson correlation = -0.0145
        - Risk model calibration/discrimination: Overall adjusted R-squared = 0.1157

- Reviewers expressed concerns regarding the decision not to include social factors despite the conceptual analyses, and for the low R-squared value and its impact on bias towards larger facilities.

Measure Developer Response

Measure #3564 Medicare Spending Per Beneficiary – Post Acute Care Measure for Home Health Agencies (Centers for Medicare & Medicaid Services)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** The Medicare Spending Per Beneficiary – Post Acute Care Measure for Home Health Agencies (MSPB-PAC HH) was developed to address the resource use domain of the Improving Medicare Post-Acute Care Transformation Act of 2014 (IMPACT Act). This resource use measure is intended to evaluate each home health (HH) agency's efficiency relative to that of the national median home health agency (HHA). Specifically, the measure assesses Medicare spending by the HHA and other healthcare providers during an MSPB-PAC HH episode. The measure reports the ratio of the payment-standardized, risk-adjusted MSPB-PAC Amount for each HHA divided by the episode-weighted median MSPB-PAC Amount across all HHAs. The MSPB-PAC Amount is the ratio of the observed episode spending to the expected episode spending, multiplied by the national average episode spending for all HHAs. The measure is calculated using two consecutive years of Medicare Fee-for-Service (FFS) claims data and was developed using calendar year (CY) 2015-2016 data. This submission is based on CY 2016-2017 data; i.e., HHA admissions from January 1, 2016 through December 31, 2017.
  Claims-based MSPB-PAC measures were developed in parallel for the HH, inpatient rehabilitation facility (IRF), long-term care hospital (LTCH), and skilled nursing facility (SNF) settings to meet the mandate of the IMPACT Act. To align with the goals of standardized assessment across all settings in PAC, these measures were conceptualized uniformly across the four settings in terms of the construction logic, the approach to risk adjustment, and measure calculation. Clinically meaningful case-mix considerations were evaluated at the level of each setting. For example, clinicians with HH experience evaluated HH claims and then gave direction on how to adjust for specific patient and case-mix characteristics.
  The MSPB-PAC HH measure was adopted by the Centers for Medicare & Medicaid Services (CMS) for the HHA Quality Reporting Program (QRP) and finalized in the CY 2017 Home Health Prospective Payment System Rate Update; Home Health Value-Based Purchasing Model; and Home Health Quality Reporting Requirements.[1] Public reporting for the measure began in Fall 2018 through the Home Health Compare website (https://www.medicare.gov/homehealthcompare/search.html) using CY 2017 data.
  Notes:
  [1] Medicare and Medicaid Programs; CY 2017 Home Health Prospective Payment System Rate Update; Home Health Value-Based Purchasing Model; and Home Health Quality Reporting Requirements. Federal Register, Vol. 81, No. 213. https://www.govinfo.gov/content/pkg/FR-2016-11-03/pdf/2016-26290.pdf
- **Type of measure:** Cost/Resource Use
- **Data source:** Assessment Data, Claims, Enrollment Data, Other
- **Level of analysis:** Facility

- **Risk-adjusted:** Yes, 124 factors, stratified by 3 risk categories, no social factors included
- **Sampling allowed**:
- **Ratings for reliability:** H-3; M-3; L-1; I-1 → Measure passes
    - Reliability testing conducted at the measure score level using the signal-to-noise analysis (Adams method) (Mean score = 0.84) and split-sample testing approach using ICCs (Mean score = 0.76); minimum of 20 episodes in a 2-year measurement period
    - Reviewers expressed some concerns with the specifications including lack of clarity which may prevent others from reliably reproducing the measure. Their concerns were regarding some of the steps to calculate the measure including exclusions, how overlapping episodes are addressed, how events are ended, and how outlier exclusions are handled
    - Reliability testing results were generally acceptable to reviewers.
- **Ratings for validity:** H-3; M-3; L-1; I-1 → Measure passes
    - Conducted empirical validity testing at the score level using multiple approaches:
        - Examined the correlation with known indicators of resource or service utilization: Hospital admissions and emergency room (ER) visits during the episode period. Compared the ratio of observed over expected spending for MSPB-PAC HH episodes with and without hospital admissions occurring in the episode period; and compared the observed over expected spending for episodes with and without ER visits.
            - Mean observed to expected ratio without ER visit=0.89
            - Mean observed to expected ratio with at least 1 ER visit =1.39
            - Without hospitalization=0.68
            - With hospitalization=2.31
        - Examined the correlation between MSPB-PAC HH scores and (i) the Discharge to Community (DTC) rates (r = -0.240) and (ii) Acute Care Hospitalization (ACH) rates (r = 0.298) for CY 2016-2017
        - Examined the correlation between MSPB-PAC HH scores with five quality measures: (all statistically significant)
            - Improvement in ambulation (#0167): r = 0.128
            - Improvement in bathing (#0174): r = 0.163
            - Improvement in bed transfer (#0175): r = 0.153
            - Improvement in management of oral medications (#0176): r = 0.141
            - Improvement in pain interfering with activity (#0177): r = 0.075
        - Risk model calibration/discrimination: Overall adjusted R-squared = 0.092
    - Reviewers expressed concerns regarding the decision not to include social factors despite the conceptual analyses, and for the low R-squared value and its impact on facility performance.
    - Reviewers also express concerns with lack of clarity around services included and the sources of variation across episodes.

Measure Developer Response

Measure #3574 Medicare Spending Per Beneficiary (MSPB) Clinician (Centers for Medicare & Medicaid Services)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description**: The MSPB Clinician measure assesses the cost to Medicare for services by a clinician and other healthcare providers during an MSPB episode, which focuses on a patient's inpatient hospitalization. The MSPB episode spans from 3 days prior to the hospital stay ("index admission") through to 30 days following discharge from that hospital. The measure includes the costs of all services during the episode window, except for a limited list of services identified as being unlikely to be influenced by the clinician's care decisions and that are considered clinically unrelated to the management of care. The episode is attributed to the clinician(s) responsible for managing the beneficiary's care during the inpatient hospitalization. The MSPB Clinician measure score is a clinician's average risk-adjusted cost across all episodes attributed to the clinician. The beneficiary populations eligible for the MSPB Clinician measure include Medicare beneficiaries enrolled in Medicare Parts A and B during the performance period.
- **Type of measure:** Cost/Resource Use
- **Data source:** Assessment Data, Claims, Enrollment Data, Other
- **Level of analysis:** Clinician : Group/Practice, Clinician : Individual
- **Risk-adjusted:** Yes, 109 factors, Stratification by 26 risk categories, No social factors included
- **Sampling allowed**: None
- **Ratings for reliability**: H-1; M-4; L-3; I-0 → Measure passes
    - Reliability testing was conducted at the measure score level using signal-to-noise (Adams method) (Mean score TIN = 0.78; TIN-NPI = 0.70) and split-sample analyses with ICC (TIN r = 0.66; TIN-NPI r = 0.60); minimum cases for TIN-NPI = 35.
    - Reviewers raised multiple concerns with attribution logic for clinicians based on specialty, winsorizing approach, and complexity of the specifications.
    - Reviewers generally found the reliability testing approach acceptable. There were concerns regarding the results for some single clinician TINs and TIN-NPIs. Specifically, those correlations and signal to noise results were less than 0.7.
- **Ratings for validity:** H-0; M-5; L-3; I-0 → Measure passes
    - Validity testing presented included systematic assessment of face validity as well as empirical validity testing.
        - Face validity approach meets NQF requirements. Results indicate experts surveyed generally agreed the measure was able to distinguish differences and accurately measure costs.
        - Developers presented multiple types of empirical validity testing:
            - Developers sought to confirm the expectation that the MSPB Clinician measure captures variation in service utilization by examining differences in risk-adjusted cost for known indicators of resource or service utilization: acute readmission and post-acute care (PAC) service utilization. They compared the ratio of observed to expected costs for MSPB Clinician episodes, with and without readmissions, and with or without PAC services utilization. Mean ratios:

  - With Readmission: 1.58
  - Without Readmission: 0.91
  - With PAC: 1.20
  - Without PAC: 0.80
- They also tested whether the measure is appropriately capturing variation in provider cost by assessing how different types of cost impact risk-adjusted measure scores. They classified costs into five clinical categories/themes and calculated the Pearson correlation between the cost of each clinical theme during the episode and the overall risk-adjusted cost for an episode

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

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

- *The MS-DRG of the index admission is included in risk adjustment
  - Reviewers generally found the face validity approach to be acceptable. However, some found the correlations performed with the cost categories to be a weak demonstration of validity.
  - Risk model discrimination/calibration analyses performed by calculating r-squared (average 0.3 across 21 MDC's) and calculated the average observed/expected cost ratio for each risk decile to demonstrate the model's prediction accuracy for high and low-cost episodes. Average O/E cost ranged 0.99-1.01 indicating good model prediction.
    - Some reviewers again expressed concern with the decision not to include social factors in the model despite conceptual and empirical analyses.
    - Other reviewers expressed concern about the adequacy of the risk model based on the correlation results indicating potential underestimation of appropriate SNF use.

Measure Developer Response

Measure #3575 Total Per Capita Cost (TPCC) (Centers for Medicare & Medicaid Services)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure

- **Description:** The Total Per Capita Cost (TPCC) measure assesses the overall cost of care delivered to a beneficiary with a focus on the primary care they receive from their provider(s). The TPCC measure score is a clinician's average risk-adjusted and specialty-adjusted cost across all beneficiary months attributed to the clinician during a one-year performance period. The measure is attributed to clinicians providing primary care management for the beneficiary, who are identified by their unique Taxpayer Identification Number and National Provider Identifier pair (TIN-NPI) and clinician groups, identified by their TIN number. Clinicians are attributed beneficiaries for one year, beginning from a combination of services indicate that a primary care relationship has begun. The resulting periods of attribution are then measured on a monthly level, assessing all Part A and Part B cost for the beneficiary for those months that occur during the performance period. The beneficiary populations eligible for the TPCC include Medicare beneficiaries enrolled in Medicare Parts A and B during the performance period.
- **Type of measure:** Cost/Resource Use
- **Data source:** Assessment Data, Claims, Enrollment Data, Other
- **Level of analysis:** Clinician : Group/Practice, Clinician : Individual
- **Risk-adjusted:** Yes, 28-133 factors, Stratification by 5 risk categories, No social factors included
- **Sampling allowed**: None
- **Ratings for reliability:** H-1; M-6; L-0; I-0 → Measure passes
    - Reliability testing was conducted at the measure score level using signal-to-noise (Adams' method) (Mean score TIN = 0.84; TIN-NPI = 0.88) and split-sample analyses with ICC (TIN r = 0.76; TIN-NPI r = 0.64); minimum cases =20.
    - Some reviewers raised concerns with the complexity of the specifications and repeating them reliably, but generally specifications were found to be acceptable.
    - Reviewers generally found the reliability testing approach and results to be appropriate and acceptable, although, one reviewer questioned the appropriate use of the Adams' method and how different variance components were obtained to calculate the scores.
- **Ratings for validity:** H-1; M-4; L-2; I-0 → Measure passes
    - Validity testing presented included systematic assessment of face validity as well as empirical validity testing.
        - Face validity approach meets NQF requirements: Results indicated, out of 15 respondents to the survey, 12 (80%) agreed that the scores from the measure as specified after comprehensive re-evaluation would provide an accurate reflection of cost
        - Developers presented multiple types of empirical validity testing:
            - Developers sought to confirm the expectation that the TPCC measure captures variation in service utilization by examining differences in mean risk- and specialty-adjusted cost for beneficiary months stratified by beneficiaries with known indicators of resource or service utilization: complications related to acute admission and post-acute care utilization. They compared the mean risk- and specialty-adjusted monthly cost for beneficiaries with and without complications related to acute admission and post-acute care utilization occurring in the measurement period.

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

Cost Driver Category	Beneficiary Mean Risk- and Specialty-Adjusted Monthly Cost

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

- They also tested whether the measure is appropriately capturing variation in provider cost by assessing how different types of cost impact risk-adjusted measure scores; they classified costs into four clinical categories/themes and calculated the Pearson correlation between the cost of each clinical theme during the episode and the overall risk- and specialty-adjusted cost for an episode.

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

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

- Attribution validity testing was demonstrated by examining the proportion of the beneficiary's Part B Evaluation and Management (E&M) codes related to primary care that are billed by the attributed TIN/TIN-NPI, to demonstrate that there is claims-based evidence that those TIN/TIN-NPIs manage their beneficiaries' ongoing care. They also conducted an impact analysis on the volume of TINs attributed the measure solely based on the services conducted by their Nurse Practitioners (NP) and/or Physician Assistants (PA), to check if TINs unlikely to manage primary care are attributed through the work of the NP and PA within their practice.

    - The mean share of beneficiary's E&M claims billed by attributed TINs or TIN-NPI is 52.8% and 45.0%, respectively.
    - 13.3% of all TINs were attributed based on services conducted by NPs and/or PAs exclusively; where 7.8% of this total come from TINs comprised of a majority of NP and/or PAs. For TINs with specialties not primary consisting of NP or PA, only 5.5% of all TINs are attributed via this method.
    - Some reviewers note the mean share of E&M claims billed to be low, and sought additional explanation of the meaning of this and how it may be accounted for in other aspects of the measure specifications.
- Risk adjustment strategy employs the HCC model which has been previously tested in the literature.
    - The R-squared reported in the December 2018 CMS Report to Congress for the CMS-HCC V22 model for community enrollees, segmented by dual eligibility and disability, range from 0.11 to 0.12. The CMS-ESRD v21 R-squared values are 0.02 and 0.11 for the dialysis new enrollee and dialysis community models, respectively.
    - Some reviewers again expressed concern with the decision not to include social factors in the model despite conceptual and empirical analyses.
    - Reviewers noted the low r-squared values.

Measure Developer Response

Measure #0369 Standardized Mortality Ratio for Dialysis Facilities (University of Michigan Kidney Epidemiology and Cost Center) (Pulled by SMP Member)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: May 15, 2008 Most Recent Endorsement Date: Dec 09, 2016]
- **Description:** Standardized mortality ratio is defined to be the ratio of the number of deaths that occur for Medicare ESRD dialysis patients treated at a particular facility to the number of deaths that would be expected given the characteristics of the dialysis facility's patients and the national norm for dialysis facilities. This measure is calculated as a ratio but can also be expressed as a rate.
  When used for public reporting, the measure calculation will be restricted to facilities with less than 3 expected deaths in the reporting year. This restriction is required to ensure patients cannot be identified due to small cell size.
- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data

- **Level of analysis:** Facility
- **Risk-adjusted:** Yes; 146 factors, Social factors included (race, ethnicity, sex)
- **Sampling allowed**: None
- **Ratings for reliability:** H-2; M-5; L-1; I-0 → Measure passes
    - Reliability testing conducted at the measure score level by calculating an inter-unit reliability (IUR) with bootstrapping; minimum 3 deaths/year to be included: IUR = 0.5, PIUR = 0.77
    - Reviewers found the reliability estimate (IUR) to be modest, but generally agreed it was acceptable.
- **Ratings for validity:** H-4; M-4; L-1; I-0 → Measure passes
    - Validity testing conducted at the measure score level by assessing the relationship of the measure to other performance measures using Spearman correlations: (all statistically significant)
        - Vascular Access: Standardized Fistula Rate (SFR): -0.08
        - Kt/V≥1.2: -0.16
        - Vascular Access: Long-term Catheter Rate: 0.07
        - Standardized Hospitalization Ratio (SHR): 0.15
        - Standardized Readmissions Ratio (SRR): 0.08
        - Standardized Transfusion Ratio (STrR): 0.16
    - Reviewers expressed concern with the low/modest correlation results, but generally found them to be acceptable as they seem to be directionally appropriate.
    - Developer also presented face validity assessment, however, because this is a maintenance measure, evaluation should rely on empirical validity testing.
    - Risk adjustment calibration/discrimination: C statistic = 0.724

Measure #1463 Standardized Hospitalization Ratio for Dialysis Facilities (University of Michigan Kidney Epidemiology and Cost Center) (Pulled by SMP Member)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: Aug 16, 2011 Most Recent Endorsement Date: Dec 08, 2016]
- **Description:** Standardized hospitalization ratio is defined to be the ratio of the number of hospital admissions that occur for Medicare ESRD dialysis patients treated at a particular facility to the number of hospitalizations that would be expected given the characteristics of the dialysis facility's patients and the national norm for dialysis facilities. This measure is calculated as a ratio but can also be expressed as a rate.
  When used for public reporting, the measure calculation will be restricted to facilities with less than 5 patient years at risk in the reporting year. This restriction is required to ensure patients cannot be identified due to small cell size.
- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes; 125 factors, 1 Social factor (sex) included
- **Sampling allowed**:

- **Ratings for reliability:** H-2; M-6; L-1; I-0 → Measure passes
    - Reliability testing conducted at the measure score level by calculating an inter-unit reliability (IUR) with bootstrapping; IUR = 0.53, PIUR = 0.75
    - Reviewers found the reliability estimate (IUR) to be modest, but generally agreed it was acceptable.
    - Reviewers expressed concern regarding the limitations in the approach to demonstrate reliability in that it is narrowly focused on identifying outliers.
- **Ratings for validity:** H-3; M-5; L-1; I-0 → Measure passes
    - Validity testing conducted at the measure score level by assessing the relationship of the measure to other performance measures using Spearman correlations: (all statistically significant)
        - Vascular Access: Standardized Fistula Rate (SFR): -0.16
        - Kt/V≥1.2: -0.23
        - Vascular Access: Long-term Catheter Rate: 0.18
        - Standardized Mortality Ratio (SHR): 0.28
        - Standardized Readmissions Ratio (SRR): 0.46
        - Standardized Transfusion Ratio (STrR): 0.42
    - Reviewers expressed concern with the low/modest correlation results, but generally found them to be acceptable as they seem to be directionally appropriate.
    - Developer also presented face validity assessment, however, because this is a maintenance measure, evaluation should rely on empirical validity testing.
    - Risk adjustment calibration/discrimination: C statistic = 0.621; Some reviewers question the decision not to include other social risk factors given the conceptual and empirical analysis presented.

Measure #2977 Hemodialysis Vascular Access: Standardized Fistula Rate (University of Michigan Kidney Epidemiology and Cost Center)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: Dec 09, 2016 Most Recent Endorsement Date: Dec 09, 2016]
- **Description:** Adjusted percentage of adult hemodialysis patient-months using an autogenous arteriovenous fistula (AVF) as the sole means of vascular access.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes, 17 factors included, no social factors included
- **Sampling allowed**: None
- **Ratings for reliability:** H-4; M-5; L-0; I-0 → Measure passes
    - Reliability testing conducted at the measure score level by calculating an inter-unit reliability (IUR) with bootstrapping; IUR = 0.755, No PIUR was provided
    - Reviewers found the reliability estimate (IUR) to be acceptable.
    - Some reviewers noted concerns with clarity of the specifications and accurately identifying comorbidities for the specifications, but generally reviewers agreed the specifications were acceptable.

        - Reviewers expressed concern regarding the limitations in the approach to demonstrate reliability in that it is narrowly focused on identifying outliers.
- **Ratings for validity:** H-1; M-7; L-1; I-0 → Measure passes
    - Validity testing was conducted at the measure score level by assessing the relationship between facility level quintiles of performance scores and the SMR and SHR using Poisson regression
        - SMR: the relative risk of mortality increased as the performance measure quintile decreased from the reference group (combined Q4 and Q5) with the highest risk in quintile 1.
            - Quintile 3, RR = 1.05 (95% CI: 1.02, 1.07; p<0.001)
            - Quintile 2, RR = 1.05 (95% CI: 1.03, 1.08; p<0.001)
            - Quintile 1, RR = 1.08 (95% CI: 1.06, 1.10; p<0.001).
        - SHR: the relative risk of hospitalization increased as the performance measure quintile decreased from the reference group (combined Q4 and Q5) with the highest risk in quintile 1.
            - Quintile 3, RR = 1.04 (95% CI: 1.03, 1.05; p<0.001)
            - Quintile 2, RR = 1.05 (95% CI: 1.05, 1.06; p<0.001)
            - Quintile 1, RR = 1.09 (95% CI: 1.08, 1.10; p<0.001)
        - Some reviewers noted the empirical validity testing results to be weak and others sought clarity on why the 4th and 5th quintiles were combined.
    - Developer also presented face validity assessment. However, because this is a maintenance measure, evaluation should rely on empirical validity testing.
    - Risk adjustment calibration/discrimination: C statistic = 0.705; Hosmer-Lemeshow statistic = 16.9, p = 0.03.
        - One reviewer noted a concern with the inclusion of sex as a social factor for risk adjustment, stating that this is inappropriate given the higher rates for unsuccessful fistula attempts among women.
        - Others noted concerns with the exclusion of social factors given the conceptual and empirical analysis presented.

Measure #2978 Hemodialysis Vascular Access: Long-term Catheter Rate (University of Michigan Kidney Epidemiology and Cost Center)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- Maintenance Measure [Original Endorsement Date: Dec 09, 2016; Most Recent Endorsement Date: Dec 09, 2016]
- **Description:** Percentage of adult hemodialysis patient-months using a catheter continuously for three months or longer for vascular access.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** No
- **Sampling allowed**: None
- **Ratings for reliability:** H-4; M-5; L-0; I-0 → Measure passes

    - Reliability testing conducted at the measure score level by calculating an inter-unit reliability (IUR) with bootstrapping; IUR = 0.76, No PIUR was provided
    - Some reviewers noted concerns with clarity of the specifications and accurately identifying comorbidities for the specifications, but generally reviewers agreed the specifications were acceptable.
    - Reviewers found the reliability estimate (IUR) to be acceptable.
- **Ratings for validity:** H-1; M-6; L-2; I-0 → Measure passes
    - Validity testing conducted at the measure score level by assessing the relationship between facility level quintiles of performance scores and the SMR and SHR using Poisson regression:
        - SMR: the relative risk of mortality showed statistically significant increases as the performance measure quintile increased from the reference group (combined Q1 and Q2) to quintile 5.
            - Quintile 3, RR = 1.03 (95% CI: 1.01, 1.05; p = 0.004)
            - Quintile 4, RR = 1.02 (95% CI: 1.00, 1.04; p = 0.063)
            - Quintile 5, RR = 1.08 (95% CI: 1.05, 1.10; p<0.001).
        - SHR: the relative risk of hospitalization increased as the performance measure quintile increased from the reference group (combined Q1 and Q2).
            - Quintile 3, RR = 1.05 (95% CI: 1.05, 1.06; p<0.001)
            - Quintile 4, RR = 1.07 (95% CI: 1.06, 1.08; p<0.001)
            - Quintile 5, RR = 1.10 (95% CI: 1.09, 1.10; p<0.001).
        - Reviewers expressed some concerns with the approach to demonstrate validity and found the results modest, but generally acceptable.
        - This measure is not risk adjusted. Some reviewers questioned the rationale for not risk adjusting but most reviewers generally found the rationale acceptable.

Measure #3565 Standardized Emergency Department Encounter Ratio (SEDR) for Dialysis Facilities (University of Michigan Kidney Epidemiology and Cost Center)

MEASURE HIGHLIGHTS

Specifications-Measure Information Form (MIF) | Testing Attachment

- New Measure
- **Description:** The Standardized Emergency Department Encounter Ratio is defined to be the ratio of the number of emergency department (ED) encounters that occur for adult Medicare ESRD dialysis patients treated at a particular facility to the number of encounters that would be expected given the characteristics of the dialysis facility's patients and the national norm for dialysis facilities. Note that in this document an "emergency department encounter" always refers to an outpatient encounter that does not end in a hospital admission. This measure is calculated as a ratio but can also be expressed as a rate.
  When used for public reporting, the measure calculation will be restricted to facilities with less than 5 patient years at risk in the reporting year. This restriction is required to ensure patients cannot be identified due to small cell size.
- **Type of measure:** Outcome

- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Yes, 86 risk factors, 1 social factor (sex) is included in the model
- **Sampling allowed**:
- **Ratings for reliability:** H-2; M-6; L-1; I-0 → Measure passes
    - Reliability testing conducted at the measure score level by calculating an inter-unit reliability (IUR) with bootstrapping; IUR = 0.62, PIUR = 0.89
    - Some reviewers noted concerns with the inclusions and clarity of specifications, but reviewers generally found the specifications acceptable.
    - Some reviewers expressed concern with the IUR result noting it as low, but acceptable, given the high PIUR.
- **Ratings for validity:** H-1; M-5; L-3; I-0 → Measure passes
    - Validity testing was conducted by stratifying facilities into two categories of SEDR: the 'better than/as expected' and 'worse than expected.' They then calculated the mean score of several quality measures:
        - Standardized Mortality Ratio (SMR)
        - Standardized fistula Rate (SFR)
        - Standardized Transfusion Ratio (STrR)
        - Percentage of Prevalent Patients Waitlisted (PPPW)
        - Standardized Hospitalization Ratio (SHR)
        - Emergency Department Visit within 30 days of discharge (ED30).
    - Then compared mean performance scores across the two strata of 'better than/as expected' 'and 'worse than expected' categories.

Table 4. Classification of SEDR and mean facility performance scores for Related Measures, 2017

		SEDR Classification		
Measure	Facilities Missing	Better than /As Expected	Worse than Expected	As Hypothesized?
SMR	310	1.00	1.08	Yes
STrR	619	0.98	1.14	Yes
SFR	395	63.49	62.12	Yes
PPPW	161	19.59	14.07	Yes
SHR	163	0.99	1.01	Yes
ED30	92	1.00	1.46	Yes

    - Developer also presented face validity assessment. However, because this is a maintenance measure, evaluation should rely on empirical validity testing.
    - Risk model calibration/discrimination: c statistic = 0.61; described as modest by reviewers

Appendix B: Additional Information Submitted by Developers for Consideration

Measure #3559 Hospital-Level, Risk-Standardized Improvement Rate in Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA) (Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE)) (Consensus Not Reached)

Reliability

- **Issue 1:** Measure specifications: Some NQF Panel members wanted clarification on the measure result calculation and definitions for "predicted," "expected" and "overall observed" improvement.
    - **Developer Response 1:** A description of the approach to measure score calculation, including a definition of each of these terms, is provided in Section 2b3.1.1 of the NQF Testing Attachment. We estimated the hospital-specific Risk-Standardized Improvement Rate (RSIR) using a hierarchical logistic regression model (hierarchical model). We calculate the hospital-specific RSIRs as the ratio of a hospital's "predicted" number of improvements to "expected" number of improvements multiplied by the overall observed improvement rate. This approach is analogous to a ratio of "observed" to "expected" that people may be familiar with. It conceptually allows for a comparison of a hospital's performance given its case-mix to an average hospital's performance with the same case-mix.

      Hospital-level RSIR Calculation =

      $$\frac{\text{Predicted Improvement}}{\text{Expected Improvement}} \times \text{Observed Overall Improvement Rate}$$

    - The expected number of cases meeting SCB improvement for each hospital (denominator) was estimated using the hospital's patient mix and the *average* hospital-specific intercept (the average intercept among all hospitals in the sample). The expected SCB improvement for each patient was calculated via the hierarchical model (HLM formula provided in NQF Testing Attachment, Section 2b3.1.1), which applies the estimated regression coefficients to the observed patient characteristics and adds the *average* hospital-specific intercept. Operationally, the expected number of cases meeting SCB improvement for each hospital was obtained by summing the expected improvement of all elective primary THA/TKA patients in the hospital.
    - The predicted number of cases meeting SCB improvement for each hospital (numerator) was estimated using its patient mix and an *estimated* hospital-specific intercept. The predicted improvement for each patient was calculated via the hierarchical model, which applies the estimated regression coefficients to the observed patient characteristics and adds the hospital-specific intercept. The predicted number of cases meeting SCB improvement for each hospital was calculated by summing the predicted improvement of all elective primary THA/TKA patients in the hospital.
    - The overall observed improvement rate is the unadjusted overall rate of SCB improvement for all patients across all hospitals.

- **Issue 2:** There was a request for clarification about how the measure accounts for patients that die between the hospital discharge and the postoperative PRO data collection period (270-365 days postoperatively), and if they are considered "lost to follow-up." Another NQF Panel member noted that excluding deaths seemed reasonable but suggested a check on death as a possible adverse event.
    - **Developer Response 2:** Patients who do not provide postoperative PROM scores (at 270 to 365 days following surgery) are not counted in the denominator or the numerator of the measure because they have incomplete PROM data. Presently, this includes patients who expire between the time of hospital discharge and the postoperative assessment window. The measure denominator is primary elective THA/TKA patients and therefore there is a lower than average competing mortality rate for this group of patients. Deaths within 30 days of the procedure are already captured in CMS' THA/TKA complications measure, with which this measure is fully harmonized. However, we will continue to work with CMS to monitor mortality and its impact on measure validity.

- **Issue 3:** An NQF Panel member noted that the HOOS, JR and KOOS, JR appeared to have been transformed from 0-100 but no specifications on the approach to transformation were provided.
    - **Developer Response 3:** Scoring of HOOS, HR and KOOS, JR are exactly as specified by the instrument developer. Stephen Lyman and colleagues[a,b] scaled the HOOS, JR and KOOS, JR to 100 points (as was done with the original HOOS and KOOS instruments), with 0 representing total hip or total knee disability and 100 representing perfect hip or knee health, respectively. Scores for the HOOS, JR and KOOS, JR were determined using Rasch-based person scores from each instruments' validation cohort. A crosswalk table provided by the authors for the HOOS, JR and KOOS, JR converts raw sum scores to the interval level measure scaled from 0 to 100. The HOOS, JR and KOOS, JR scores were derived from the responses to full HOOS surveys from both registries.
- **Issue 4:** An NQF Panel member noted that the interval over which the "change" in score appears to have been estimated (90-0 days prior to surgery and 270-365 days following surgery) is quite wide and could vary for an individual patient by as much as 6 months.
    - **Developer Response 4:** The Technical Expert Panel (TEP) considered both data assessment timeframes very carefully. When considering the preoperative assessment window, the TEP believed that elective primary THA and TKA candidates were unlikely to have significant changes in preoperative PROM scores within 90 days of surgery. In addition, they indicated that the additional time to collect data would increase response rates and likely better represent stable and complete recovery from either procedure. Likewise, the postoperative assessment window was considered very carefully. After considerable input from TEP members, public comments, a thorough literature review, and a review of registry experiences, we defined the postoperative PROM data collection timeframe to between 270 days and 365 days. The TEP concurred with this recommendation. This timeframe allows for full recovery from both THA and TKA and increases opportunity for PRO response.

---

[a] Lyman S, Lee Y-Y, Franklin PD, Li W, Mayman DJ, Padgett DE. Validation of the HOOS, JR: A Short-form Hip Replacement Survey. *Clinical Orthopaedics and Related Research®.* 2016;474(6):1472-1482.

[b] Lyman S, Lee Y-Y, Franklin PD, Li W, Cross MB, Padgett DE. Validation of the KOOS, JR: A Short-form Knee Arthroplasty Outcomes Survey. *Clinical Orthopaedics and Related Research®.* 2016;474(6):1461-1471.

- **Issue 5:** An NQF Panel member noted concern about attributing changes in joint function to the hospital (versus care such as rehabilitation services) with a follow-up interval of nine months to one year following surgery.
    - **Developer Response 5:** The goal of this hospital-level PRO-PM is to capture the full spectrum of care to incentivize collaboration and shared responsibility for improving patients' health and reducing the burden of their disease. As per our response on the NQF Evidence Form, Section 1a.2.**,** we note that optimal clinical outcomes for patients undergoing an elective primary THA or TKA depend not just on the surgeon performing the procedure, but also on the entirety of the team's efforts in the care of the patient, care coordination across provider groups and specialties; and the patients' engagement in their recovery. Even the best surgeon will not get outstanding results if there are gaps in the quality of care provided by others caring for the patient before, during, and/or after surgery. Further, the hospital has significant influence over discharge and rehabilitation planning for its surgical patients.
- **Issue 6:** An NQF Panel member noted that this appears to be a composite measure, but that NQF form does not appear to have been completed.
    - **Developer Response 6:** This is not a composite measure. The outcome measure is not a composite of a THA PRO-PM and a TKA PRO-PM. Instead, the cohort for this measure consists patients undergoing an elective primary total hip or total knee arthroplasty, and outcomes for patients in the cohort are determined using a single risk model.
- **Issue 7:** Some NQF Panel members had questions about the 25 case volume threshold—what the threshold was based on, what happens to a facility that falls below the 25-case recommendation, if facilities without 25 cases would be excluded from the measure (and should be identified as an exclusion), and if excluded, whether it would create an incentive for them to not complete data.
    - **Developer Response 7:** A 25 case volume threshold is consistent with volume thresholds used for public reporting of claims-based measures with which this measure was intentionally harmonized. It is <u>not</u> a measure exclusion; rather, the recommendation is that hospitals that perform fewer than 25 elective primary THA or TKA procedures during the measurement period or have complete PRO data on fewer than 25 THA or TKA procedures during the measurement period not be included in public reporting of the measure. This recommendation is made to address concerns about the reliability of measure results for hospitals with a small number of procedures and/or procedures with PRO data. And the aggregate number of elective primary THA/TKA procedures conducted among hospitals performing fewer than 25 of these procedures is small; while 33% of hospitals conducted fewer than 25 elective primary THA and TKA procedures from July 1, 2016 to June 30, 2017, the procedures performed at these hospitals represented just 3.14% (11,175 of 333,850) of the total number of elective primary THA and TKA procedures performed across all hospitals.
    - It is expected that hospitals with fewer than 25 procedures total or fewer than 25 procedures with complete PRO data would still receive hospital specific reports, informing them of measure results, but that a Risk-Standardized Improvement Rate (RSIR) for these hospitals would not be publicly reported. Hospital-specific reports provided to these hospitals might also positively impact the collection of PRO data even if an RSIR was not publicly reported.
- **Issue 8:** An NQF Panel member had a clarification question about the data used for reliability testing.
    - **Developer Response 8:** The Combined Dataset consists of the hospitals in both the Development and Validation Datasets combined that have at least 25 elective primary

THA/TKA Patients with PRO Data. These 123 hospitals make up the Combined Dataset used for reliability and validity testing. This is consistent with the measure as specified with a 25-case volume threshold.

- **Issue 9:** Concern was expressed about data element reliability testing for "critical data elements" other than the HOOS, JR and the KOOS, JR. Data elements of concern were noted to be those that "make-up the denominator," the two additional PRO tools used in the risk model, and, additional risk factors, including the clinical characteristics based on coding (e.g. liver disease, severe infection).
    - **Developer Response 9:** The codes used to define the measure cohort (denominator) are harmonized with CMS' publicly reported, NQF-endorsed hospital-level THA/TKA complications measure. This measure has been in public reporting since 2013 and undergoes annual updates through independent clinical review by orthopedic coding experts to ensure the measure methodology reflects current clinical and coding practice. In addition, we only use data elements in claims that have both face validity and reliability. We do not use fields that are inconsistently coded across providers. We only use fields that are consequential for payment and which are audited. We identify these variables through empiric analyses and our understanding of CMS auditing and billing policies and do not use variables which do not meet this standard. 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 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. While we have not performed medical record chart review validation of this measure risk model, multiple CMS claims-based measure risk models have been validated using chart review, a few of them cited here.[c,d,e]
    - The risk variables included in this risk model were defined over several years through multiple, iterative steps that pulled in stakeholder input on feasibility, clinical capture, accuracy, reproducibility and clinical face validity. These steps included surveying orthopedic practices regarding the feasibility, uniformity and reliability of risk variables identified by clinical experts and published literature; a consensus summit by orthopedic specialty societies to narrow and prioritize clinical risk variables for prospective collection as part of the CJR model – these recommendations were adopted in toto by CMS; additional clinical and empiric evaluation in CJR data; and by TEP approval.
    - Patients and the TEP were engaged throughout the measure development process. The TEP was thoroughly engaged in the selection of risk variables for inclusion in the risk model, providing input on the importance and feasibility of each variable. Both the TEP

---

[c] Krumholz H, Normand S, Keenan P, et al. Hospital 30-Day Pneumonia Readmission Measure Methodology [Internet]. Yale New Haven Health Services Corporation/ Center for Outcomes Research and Evaluation;2008. Available at: http://www.qualitynet.org/dcs/ContentServer?c1/4Page&pagename%QnetPublic%2FPage%2FQnetTier3&cid1/4121906985841.

[d] Krumholz H, Normand S, Bratzler D, et al. Risk-Adjustment Methodology for Hospital Monitoring/Surveillance and Public Reporting Supplement#1: 30-Day Mortality Model for Pneumonia [Internet]. Yale University;2006. Available at: http://www.qualitynet.org/dcs/ContentServer?c1/4Page&pagename%QnetPublic%2FPage%2FQnetTier3&cid1/41163010421830.

[e] Keenan PS, Normand SLT, Lin Z, et al. An Administrative Claims Measure Suitable for Profiling Hospital Performance on the Basis of 30-Day All-Cause Readmission Rates Among Patients With Heart Failure. *Cardiovascular Quality and Outcomes*. 2008;1(1):29-37.

and the Patient Working Group provided detailed input on the measure outcome definition.

- **Issue 10:** An NQF Panel member noted that developers reported that HOOS JR was not tested for reliability because the HOOS was tested several times, and do not state it was tested here.
    - **Developer Response 10:** As required by NQF, reliability and validity of the HOOS, JR and KOOS, JR are provided in detail in the NQF Testing Attachment. Section 2a2.3 clarifies that internal consistency reliability using the Person Separation Index (PSI) was assessed for the HOOS, JR, as was a principal component analysis, conducted on the standardized residuals and indicating that the six HOOS, JR items existed in a single dimension. Only test-retest reliability for the HOOS, JR was dependent on results from assessment of the HOOS domains from which the HOOS, JR pain and functioning questions were drawn.
- **Issue 11:** An NQF Panel member stated that "Reliability testing appears to have been done at the patient not hospital level (the unit of comparison of the measure)" and later noted that an ICC comparing hospital level results appears not to have been performed.
    - **Developer Response 11:** Per NQF Guidance (for example, see page 2, Note 10 of the NQF Testing Attachment Form as well as check-off box, Section 2a2.1 suggesting the use of signal-to-noise analysis for measure score validity), we conducted hospital-level reliability testing with signal-to-noise analysis for comparison of hospital-level measure scores. Results are provided in Section 2a2.3 of the NQF Testing Attachment.
- **Issue 12:** An NQF Panel member voiced concern about proxy assessment, noting that it "is unorthodox and can add significant noise."
    - **Developer Response 12:** As this is a measure of elective procedures, proxy assessments are uncommon (in our data, of the 81% of data submissions with respondent identified, only 8.8% were identified as surrogate responses) but we chose to include these patients in order to ensure they were being measured. We will advise CMS to continue to examine these patients in reevaluation.
- **Issue 13:** Two NQF Panel members voiced concern about missing data, and that the only complete data were analyzed without accounting for what is likely "fairly extensive missingness." One of these members noted concern that missing surveys were accounted for but that missing responses within the survey were not.
    - **Developer Response 13:** We provide a detailed accounting in the NQF Testing Attachment in Sections 2b6.1 through 2b6.3 of our approach to PRO non-response (including elective primary THA and TKA patients with no PRO data and patients missing either preoperative or postoperative data or missing or out-of-range values on PRO data submitted). Due to the voluntary nature of PRO survey data and because PRO data are unlikely to be missing at random, we understand that accounting for potential non-response bias is important for this measure. Since bias may be introduced by systematic differences between responders such as patients with different social risk, we included social risk factors and race in the propensity score models used to create stabilized inverse probability weights to address potential non-response bias.
    - On Section S16 of the NQF Submission Form, we note the importance of high response rates for measuring hospital quality with PROs: "High response rates allow PRO-PMs to best represent hospital quality performance. Hospitals and physicians incorporating PRO data collection into clinical workflows are likely to reap considerably higher response rates. Strong leadership support within the hospital, flexibility in rearranging clinical workflows to accommodate PRO data collection, accessibility of PRO data in real-time to inform clinical decision making can all increase staff investment in the value of PROs in

improving care and quality, and PRO data used for clinical decisions can increase patient investment."

- Regarding missing responses within the survey data, in Section 1.7 of the NQF Testing Attachment, we state that only PROs with complete data are used in measure development and testing. [Complete PRO data is defined as the submission of preoperative patient-reported outcome measure (PROM) and risk variable data with no missing or out-of-range values for required data elements and that could be matched to postoperative PROM data with no missing or out-of-range values, for an elective primary THA/TKA procedure identified in claims data for the measurement period.] For the voluntary data collection, missing data are better addressed through accounting for non-response bias than through data imputation. The TEP supported this approach.

Validity

- **Issue 1:** An NQF Panel member suggested that the exclusion of staged procedures might eliminate up to 43% of procedures, and that the measure name should include "from unstaged procedures."
  - **Developer Response 1:** Please note that, globally, the prevalence of staged THA/TKA procedures that are not simultaneous and occur within 1 year of each other is roughly 7%.[f,g] We are happy to clarify this in the measure name if the committee feel this is required.
  - Among hospitals submitting PRO data, 7.06% of THA and TKA procedures were staged procedures during the measurement period (two or more procedures occurring during the measurement period in distinct hospitalizations).
  - As we note in Section 2b2.2 of the NQF Testing Attachment, 491 (4.17%) of patients with complete PRO and risk variable data had staged procedures during the measurement period. Across hospitals, the mean proportion of procedures excluded from the analysis was 3.84% (SD 5.69), and the median proportion was 2.11%.
- **Issue 2:** An NQF Panel member noted concern that data were not provided on how the excluded patients impact the performance measure scores.
  - **Developer Response 2:** Because the assessment of the measure outcome is unclear in patients with staged procedures – that is, it is hard to clarify the impact of the index procedure on the PRO result – we did not include staged procedures in the measure score. Our clinical consultants and Technical Expert Panel agreed with this exclusion. We are happy to recommend to CMS that staged procedures be reexamined during reevaluation. Because this exclusion is based on the inability to appropriately attribute the outcome to the index procedure, we are uncertain how to interpret the results requested by the Panel member.
- **Issue 3:** Concern was expressed about the 25-case volume recommendation: that it is not identified as an exclusion, that this represents 52% of hospitals in the denominator, and that not

[f] Stefánsdóttir A, Lidgren L, Robertsson O. Higher early mortality with simultaneous rather than staged bilateral TKAs: results from the Swedish Knee Arthroplasty Register. *Clin Orthop Relat Res*. 2008;466(12):3066–3070. doi:10.1007/s11999-008-0404-3

[g] Garland A, Rolfson O, Garellick G, Kärrholm J, Hailer NP. Early postoperative mortality after simultaneous or staged bilateral primary total hip arthroplasty: an observational register study from the Swedish Hip Arthroplasty Register [published correction appears in BMC Musculoskelet Disord. 2015;16:263]. *BMC Musculoskelet Disord*. 2015;16:77. Published 2015 Apr 8. doi:10.1186/s12891-015-0535-0

considering or testing hospitals that fall below this threshold is a major potential threat to the measure's validity unless the denominator is redefined as suggested above.

- **Developer Response 3:** As noted in Issue #3 for Reliability above, a 25-case volume threshold is consistent with volume thresholds used for public reporting of claims-based measures with which this measure was intentionally harmonized. It is not a measure exclusion; the recommendation is that hospitals that perform fewer than 25 elective primary THA or TKA procedures during the measurement period or have complete PRO data on fewer than 25 THA or TKA procedures during the measurement period not be included in public reporting of the measure. However, these hospitals will receive confidential measure results.

- **Issue 4:** Another NQF Panel member: The model was developed including cases from hospitals not used for reliability, validity and missing data testing, i.e., hospitals with low caseloads (n<25) not recommended for this measure. Did the developers do a sensitivity test to assess the impact of excluding these hospitals from the risk-adjustment development sample on the risk-adjustment model?
  - **Developer Response 4:** The risk model was developed using all cases in the Development Dataset and validated using all cases in the Validation Dataset. The recommendation for a 25-case volume threshold is for public reporting, and therefore reliability and validity analyses were conducted on hospitals with at least 25 elective primary THA and TKA procedures with PRO data. Including all the THA or TKA procedures for the risk model development will maximize the available information and is a commonly accepted approach.
  - Analysis to address non-response included all THA and TKA procedures conducted at all 238 hospitals. The hospitals with 25 or more procedures are reported with weighting for non-response, as per the recommendation for a 25-case volume threshold for public reporting.
- **Issue 5:** An NQF Panel requested clarity for the data provided in T.11 and whether there are meaningful differences between hospitals in the top quartile.
  - **Developer Response 5:** Table 11 indicates a range of Risk-Standardized Improvement Rates (RSIRs) from the 75th to the 100th percentile of hospitals of 72.51% to 86.84%. These RSIRs indicate the risk-standardized proportion of patients achieving substantial clinical benefit improvement following elective primary THA or TKA. The 14.3 percentage points representing this range represent a meaningful difference in the proportion of patients experiencing substantial clinical benefit improvement.
- **Issue 6:** An NQF Panel member asked if the impact of IPW on hospital ratings was assessed by conducting a sensitivity analyses?
  - **Developer Response 6:** In the NQF Testing Attachment in Section 2b6.2, Table 14 we provide a comparison of the mean and distribution of hospital RSIRs with and without stabilized inverse probability weighting. As we note in interpretation of results in Section 2b6.3, this comparison reveals only a small impact on the measure results of adjusting for potential non-response. However, we expect that non-response bias will be a factor for the THA/TKA PRO-PM measure, due to associations with non-response including socioeconomic status and health status. We therefore retained response bias adjustment for the measure results.
- **Issue 7:** Two NQF Panel member**s** expressed additional concern about the extent of missing data and the subsequent threat to measure validity (also noted under Reliability, Issue 13 above). One Panel member noted their belief that the proposed solution (stabilized inverse probability

weighting) assumes data missing at random and suggested that requiring near-complete data rather than rely on proxies and statistical modeling was the only good solution.

- **Developer Response 7:** As noted in our response to Issue 13 under Reliability (above), due to the voluntary nature of PRO survey data and because PRO data are unlikely to be missing completely at random, we understand that accounting for potential non-response bias is important for this measure. Since bias may be introduced by systematic differences between responders such as patients with different social risk, we included social risk factors and race in the propensity score models used to create stabilized inverse probability weights to address potential non-response bias.
- Stabilized inverse probability weighting, calculated using propensity model, does not assume that data are missing completely at random; rather, that particular patient groups have different response rates that are accounted for in the weighted model.
- On Section S16 of the NQF Submission Form, we note the importance of high response rates for measuring hospital quality with PROs: "High response rates allow PRO-PMs to best represent hospital quality performance. Hospitals and physicians incorporating PRO data collection into clinical workflows are likely to reap considerably higher response rates. Strong leadership support within the hospital, flexibility in rearranging clinical workflows to accommodate PRO data collection, accessibility of PRO data in real-time to inform clinical decision making can all increase staff investment in the value of PROs in improving care and quality, and PRO data used for clinical decisions can increase patient investment."

- **Issue 8:** An NQF Panel member asked why the overall observed improvement rate would be used both in the development of the HLM as the dependent variable, and then again in the calculation of the RSIR?
  - **Developer Response 8:** The overall observed improvement rate is used in the calculation of the hospital-level RSIR (as noted in Issue #1 under the "Reliability" heading above, it is a constant to assist the interpretation of RSIR and has no material impact on RSIR) but is not the dependent variable of the HLM model. The dependent variable for this model is a patient-level outcome, identifying the individual patient's improvement.
- **Issue 9:** An NQF Panel member asked how health literacy how will be measured in practice.
  - **Developer Response 9:** In Section 2b3.1.1 of the NQF Testing Attachment, we list the variables included in the final risk model and note that Health Literacy is assessed by response to the Single Item Literacy Screener questionnaire, which asks about "Comfort Filling Out Medical Forms by Yourself"). The response options are noted in the Data Dictionary in Row 5 of the "Risk Variables with PRO Data" tab: 0 = Not at all, 1 = A little bit, 2 = Somewhat, 3 = Quite a bit, 4 = Extremely.
- **Issue 10:** Concern was expressed about data element validity testing, that published validity data from the HOOS, JR and KOOS, JR were provided, but testing for other critical data elements was not provided.
  - **Developer Response 10:** Please see the detailed response to Issue 9 under Reliability above.
- **Issue 11:** A few NQF Panel members noted concern about ceiling effects of the HOOS and KOOS. One member noted that recent publications have supported the use of other non-condition specific measures (e.g. PROMIS physical function) as valid alternatives for future consideration.
  - **Developer Response 11:** Thank you for this input. We will be sure that CMS and the measure reevaluation contractor are provided this suggestion.

- **Issue 12:** There was a question about the logic in selecting a single threshold for SCB (by THA/TKA). It was noted that there is "a wealth of published literature on the dependency of clinically important improvement thresholds on initial scores." It was suggested that patients with worse initial scores would need to see greater improvement to reach "clinically important improvement thresholds" than patients with higher initial scores. Concern was raised about the Measure Developer's statement that the SCB outcome allows patients with poor baseline PRO scores to improve, that some risk variables that might be traditionally considered as predictors of worse outcomes are positively associated with achieving a SCB, and that this biases the measure and may not meet a patient's expectations of improvement. Also, concern was expressed that this approach would penalize providers with higher performing patients at admission.
    - **Developer Response 12:** With strong TEP support, this PRO-PM measures improvement with a threshold for the HOOS, JR and for the KOOS, JR tested by Stephen Lyman and colleagues[h] (developers of the HOOS, JR and KOOS, JR) and identified using an anchor-based question to assess substantial clinical benefit (SCB, 22 points for HOOS, JR and 20 points for KOOS, JR). This improvement threshold approach to the outcome was preferred over alternatives (averaging change among patients, measuring a postoperative average or minimum state, or a combined approach of improvement and postoperative state). An improvement threshold approach was preferred for the following reasons:
        - It measures improvement only and discourages surgeons from performing THA/TKA procedures on patients with milder symptoms, as patients with high preoperative PROM scores cannot statistically meet the improvement threshold;
        - It equally rewards hospitals performing THA/TKA on patients with moderate and severe symptoms, as it does not define an "end state" that patients must achieve, only substantive improvement from where they started;
        - Avoids creating what is known as a ceiling effect, where many patients can meet the outcome criteria and decreases the ability of the measure to identify performance variation; and
        - It has less risk of unintended consequences. Specifically, we were concerned that requiring patients to meet a postoperative minimum symptom state would encourage hospitals and their surgeons to avoid offering THA/TKA surgery to anyone with severe pain and/or limited function, the people most in need of surgery.
    - Some risk variables that might be traditionally considered as predictors of worse outcomes are positively associated with achieving a SCB because patients with more severe symptoms at baseline have more opportunity for improvement. Patients on our TEP and on our Patient Working Group supported this improvement threshold.
    - The TEP supported a lower opportunity for patients with high scores preoperatively, indicating that mild symptoms, to reach substantial clinical benefit improvement. TEP members were in favor of a measure that dis-incentivized inappropriate surgery and clinicians performing major elective surgery on patients will little opportunity for benefit.

[h] Lyman S and Lee YY. What are the minimal and substantial improvements in the HOOS and KOOS and JR versions after total joint replacement? *Clinical Orthopaedics and Related Research®.* 2018;467(12):2432-2441.

- **Issue 13:** An NQF Panel member voiced concerns with the lack of adjustment for non-English speakers, given that the KOOS, Jr. and HOOS, Jr. are only offered in English.
    - **Developer Response 13:** We do not have available to us a variable representing primary or spoken language, preventing risk adjustment consideration. Active efforts to make the HOOS, JR and KOOS, JR available in additional languages are ongoing. We will recommend to CMS that it considers collecting preferred language status for possible risk adjustment.
- **Issue 14:** An NQF Panel member noted concern that the only social risk factor included is health literacy.
    - **Developer Response 14:** Health literacy is a potent predictor and associated with a range of social risk factors. In addition, as noted in the NQF Testing Attachment, Section 2b3.4b, the results of the social risk factor testing did not provide evidence of significant differences in measure results. However, we did find that social risk factors were significantly associated with response and therefore, we included social risk in our non-response adjustment of the measure. As this measure assesses patients undergoing an elective procedure where known disparities exist, we will recommend CMS continues to assess the impact of social risk for this measure over time.
- **Issue 15:** An NQF Panel member noted large differences between NQF #1550 groups on data element, and that few patients appear to report substantial clinical improvement, noting that this could be because the bar is set too high, or ceiling effects of the measures, or both.
    - **Developer Response 15:** It appears that this Panel member is referring to Figure 1 in the NQF Testing Attachment when referring to NQF#1550 (the THA/TKA Complications measure used for Empiric Measure Score validity assessment). The comment regarding "few patients appear to report substantial clinical improvement" is not understood. This figure shows that hospitals with worse than the national average complication rates have a median RSIR just above 50%, whereas hospitals at the national average complication rates have a median RSIR near 65% and hospitals with better than national average complications rates has a median RSIR at approximately 70%. Table 14 of the NQF Testing Attachment notes that the risk-standardized mean RSIR for hospitals is 60%.
- **Issue 16:** An NQF Panel member noted that the THA/TKA PRO-PM RSIR with the hospital risk standardized complication rate (NQF: 1550) displayed box plots with evidence of considerable validity in results at the mean. A plot of the association of pass/fail on each measure at the hospital level would have been helpful.
    - **Developer Response 16:** As outcome measures are often reported as point estimates with uncertainty ranges reflecting the statistical uncertainty inherent in outcome measurement, we felt it better to represent the statistical uncertainty that CMS reports for NQF 1550 than to report validity using only the point estimate and without acknowledging the statistical uncertainty. As CMS has not yet indicated it plans for reporting RSIRs, we did not apply any calculation of statistical uncertainty to the RSIRs.

Other General Comments

- **Issue 1:** NQF Panel Member #1 state that specifications of the measure identified that it was both risk-stratified and risk-adjusted, and that the difference between these two terms were not provided.
    - **Developer Response 1:** This statement is not consistent with the information we provided. The measure is identified in different sections as risk-standardized (or risk-

adjusted) but not risk-stratified. The terms risk-standardized and risk-adjusted both signify that the measure is risk-adjusted.

- **Issue 2:** NQF Panel Member #8 noted that the variables listed with measure specifications in S5 for risk adjustment did not match those listed in 2b3.
    - **Developer Response 2:** In S5, we identify the data elements listed as those used to define the numerator and for risk adjustment that are collected with PROM data; this is not intended to be a complete list of risk variables in the risk adjustment model.
- **Issue 3:** There was a request for explanation of why multiple risk adjustment variables in the risk-adjustment model were included that where not significant?
    - **Developer Response 3:** When building claims-based models, we have previously used the strength of association between the risk variable and the measure outcome to empirically guide risk variable selection. When expert input deems it appropriate, we force in additional risk variables, such as those that indicate frailty, that might have an important influence on the measure outcome and yet might not be selected for the model based purely on statistical considerations. In this way, our risk models always reflect both empirical data and clinical input. This approach has produced robust risk models that have been repeatedly and successfully validated against medical record data. For this measure, we applied the same principles, but recognize that PRO-PM development, particularly that based upon a voluntary data sample, may require a greater reliance on clinical input to select risk variables than traditional claims-based outcome measures. Therefore, for this measure, we conducted analyses to evaluate two approaches to risk model development for each PROM outcome – one used a purely data-driven approach (referred to as the empirically derived model) and another used candidate risk variable selection based on empirical findings in the literature, review of data-driven risk factors, and iterative TEP and clinical expert input and ranking of importance and feasibility of risk variables for a THA/ TKA PRO-PM (referred to as a clinically derived model). We identified an extensive list of risk variables for consideration in the development of the risk model(s), through a systematic literature review and environmental scan, as well as from orthopedists surveyed about what risk variables they consider important in predicting THA/TKA outcomes. In consultation with the Technical Working Group and the TEP and through detailed public comments from specialty societies, we focused on candidate risk-adjustment variables of interest that were clinically relevant and had an evidence-based relationship with clinical outcomes following elective primary THA or TKA. Likewise, we considered several potential data sources, including administrative claims, registry- or clinician-provided data, and patient-reported sources. In addition to clinical risk variables that have been collected de novo and evaluated for inclusion in the final measure risk model, all diagnostic codes from administrative claims during the 12 months prior to the THA/TKA procedure or secondary diagnosis codes during the index admission except those associated with potential complications during the index admission were evaluated for possible inclusion in the risk model. Recognizing the thorough vetting of risk variables for this risk model, we determined to keep variables in the model that may not reach statistical significance in our data with an understanding that our sample may be limited and that this risk model will be implemented more broadly.
- **Issue 4**: An NQF Panel member suggested that the creation of the HOOS, JR and KOOS, JR instruments were never discussed, and that they would have liked evidence of the content coverage (content validity) for each measure.

- **Developer Response 4:** The HOOS, JR and KOOS, JR were developed at the Hospital for Special Surgery by Stephen Lyman and colleagues. The instruments are non-proprietary, free to use, and were validated in 2016. Reliability and validity testing conducted for these PROM surveys is reported in manuscripts on the validation of these instruments[i],[j] and noted in the NQF Testing Attachment form, Sections 2a2.2 and 2a3.3.

Measure #0715: Standardized adverse event ratio for congenital cardiac catheterization (Boston Children's Hospital - Center of Excellence for Pediatric Quality Measurement)

Reliability

- **Issue 1: Numerator**
  - **Concern: Clarity on Definition of Major Adverse Events**
    - We have defined major adverse catheterization-related events (severity level 4 and 5) based on specific previously published International Pediatric and Congenital Cardiac Code (IPCCC) classification schema, such that level 4 AEs result in a change in the patient's clinical condition which would be life-threatening if not treated and which require intensive medical therapy and/or major invasive transcatheter or urgent/emergent surgical intervention to treat the condition. These conditions may also result in the need for unplanned cardiopulmonary support to prevent a catastrophic event from occurring. Some examples include: a major life-threatening vascular injury which results in cardiopulmonary collapse, need for urgent blood product administration, and/or a major invasive procedure to successfully treat the condition; any event requiring cardiopulmonary resuscitation (CPR); emergent surgical intervention due to device or stent embolization; and unanticipated intubation in the setting of circulatory collapse or acute respiratory failure. Level 5 events are catastrophic complications resulting in subsequent death of the patient due to procedural complication.
    - While the adverse events themselves may be heterogeneous, the clinical responses to major adverse events are often very similar - such as an escalation in hospital-level care.
    - The goal of this metric is to develop a single ratio which quantifies whether a hospital is experiencing a higher or lower rate of adverse events than would be expected given the complexity of their patient population and the procedures they perform.
    - Changing the outcome variable to major adverse events (severity level 4,5) for CHARM II from the previous outcome clinically significant adverse events (severity level 3,4,5) for CHARM I, the previously endorsed NQF metric, will be less susceptible to recording bias.
    - Reviewers raised concerns around heterogeneity of the chosen *major* adverse event outcome, but this is of greater concern for *clinically important* high severity adverse events (severity level 3,4,5). We do not have data to prove this, as the level 3 events were recorded as reliably as level 4 and 5 events in the audit (clarification in issue #4). In general use, there is face validity in the assumption that recording bias will be less of a concern for these unequivocal life threatening major adverse events as compared to "clinically" important adverse events.
    - Patients may experience multiple severe adverse events which may and often does

[i] Lyman S, et al. Validation of the HOOS, JR (see footnote 1, page 2)
[j] Lyman S, et al. Validation of the KOOS, JR (see footnote 2, page 2)

occur due to a singular instigating event or due to poor clinical status of the patient.
- The outcome for the metric is in fact ***binary*** and is the occurrence of "any" Major Adverse Event. We apologize that this was miscommunicated with our use of "a" major adverse event in the submission.

♦ **Issue 2: Denominator Definition Clarification and Details on Exclusions**

- **Clarity on Definition of Congenital Heart Disease Catheterization**
  - Congenital Heart Disease Cardiac Catheterization is primarily defined by, but not limited to, catheter- based interventions aiming to diagnose and/or treat conditions related to congenital malformations in heart structure, and require expert treatment and diagnosis at tertiary cardiac centers. These cases may also include acquired cardiac conditions, generally in young patients, who require specialized diagnostic or transcatheter therapies.
  - Institutions included in this measure are centers which provide expert care to this unique patient population of infants, children and adults with cardiac disease who are distinctly different from adults requiring cardiac catheterization for coronary artery disease.

- **Clarity on Inclusion: Age**
  - The CHARM II model used for the SAER metric was developed in a data set including all eligible cases (n=23212) with no age exclusions.
  - The model was derived in a 75% random sample of the cohort and tested in the remaining 25%. Patients age 18 or younger comprised 88% of the cohort and patients 19 or older comprised the remaining 12%.
  - To clarify for the reviewers, we also validated the performance of the CHARM II model in a data set consisting entirely of cases in patients less than or equal to 18 years of age (n=20,502). This was done to assess model performance, and thus generalizability of the SAER metric, for institutions which only provide care for the pediatric population. Model discrimination and calibration in the pediatric cohort was equivalent to that in the full cohort.

- **Clarity on Exclusions: Hospitals Ineligible for Metric**
  - All hospitals in the testing and validation cohorts performed more than 50 cases per year. While a minimum case volume is a recommendation for metric interpretability, we do not have data on the impact of excluding hospitals with less than 50 cases. The metric will be less meaningful for institutions with very small population sizes, as this results in a wide confidence interval around the SAER.

- **Clarity on Exclusions: Cases Ineligible for Registry**
  - The target population for this metric is patients undergoing diagnostic and interventional procedures on congenital malformations of the heart. Case types are defined by published IPCCC nomenclature for procedure types and chosen from a list of options by centers performing cardiac catheterization in this population.
  - Pericardiocentesis and thoracentesis are draining procedures that do not require catheter access but may be performed in the catheterization lab to utilize concurrent radiographic guidance. Because they do not require vascular catheter access and are not procedures performed on the heart or surrounding vessels, these cases are excluded.
  - Additionally, catheterization cases for the purpose of evaluating and treating rhythm disturbances (electrophysiology cases) were not eligible for the registry,

and thus no data on these cases were collected.

- **Clarity on Exclusions: Procedures Eligible for Registry but Excluded from the Denominator – Conditions/Procedures Comprising Exclusions (S.8, S.9)**
    - The following is a list of the types of cases that were not assigned to a CHARM II risk category:
        - Other defect or vascular closure
        - Other transcatheter valve procedure
        - Tricuspid valvotomy
        - Aorta (other) dilation and/or stent
        - Systemic artery (not aorta) dilation and/or stent
        - Systemic pulmonary collateral dilation and/or stent
        - Systemic vein dilation and/or stent
        - Other angioplasty and/or stent
        - Fenestration dilation and/or stent
        - Foreign body removal
        - Coronary fistula closure
        - Paravalvar leak closure
    - These cases do not represent missing data. Patient, procedural, and outcome information was collected for these cases. However, these cases were excluded from the denominator because only case types with a corresponding assignment to a designated procedure type risk category are eligible.
    - The case types above were not assigned to a procedure type riskcategory according to expert opinion for the following reasons:
        - The primary reason for exclusion is the heterogeneity of expected outcomes at the case level.
        - Some of these case types were excluded because the location specified could include subcategory intervention locations, such systemic artery, which may include both renal and iliac vessel interventions with different risk profiles.
        - Additionally, at a specified anatomic location, the indications for interventions may be so different that the expected outcomes vary widely.
        - Some cases have heterogeneity in the complexity of the case for the same intervention type, for example coronary fistulas, paravalvar leaks, and foreign body removals.
        - Others may represent novel, rarely performed cases, or procedures in unusual anatomic
        - locations designated as “other” with limited procedural and outcome information.
    - **Appendix 1** is a table summarizing patient characteristics and adverse event rates in cases without a case type risk category designation, which were excluded from analysis, compared to the analysis cohort. This is supplemental data to the summary response in the submission2b2.2.
    - Reviewer comments suggested sensitivity testing on the impact of excluding these cases. In response we have provided results of the SAER metric at the institution level with and without the excluded cases (see **Table 1**, below). For testing purposes, the cases without a CHARM II case type designation that were excluded from our analysis were added back into the analysis cohort and assigned to a sixth CHARM II risk category (separate from defined risk categories 1-5). The model for the outcome “any level 4/5 adverse event” was fitted with this expanded set of risk categories, hemodynamic score, and age group, and standardized adverse event ratios were re-estimated for the 13 institutions in the

study cohort. Differences in the revised SAER were fairly minor, and well within the given 95% Confidence Intervals, as shown in the table below. The Spearman rank correlation between the original SAERs and the revised SAERs, which incorporate the previously excluded cases, is r = 0.97. This indicates that the exclusion of these cases from our analysis cohort does not impact the validity of our model.

Table 1: SAER by Institution and Revised SAER Including the Excluded Cases

Institution	SAER	95% Confidence Interval	Revised SAER
1	0.33	(0.07, 0.98)	0.36
2	0.34	(0.15, 0.68)	0.35
3	0.47	(0.09, 1.36)	0.55
4	0.59	(0.16, 1.49)	0.68
5	0.74	(0.37, 1.33)	0.76
6	0.93	(0.60, 1.39)	0.89
7	0.99	(0.64, 1.47)	1.07
8	1.01	(0.81, 1.23)	0.98
9	1.07	(0.74, 1.52)	1.04
10	1.19	(0.86, 1.62)	1.17
11	1.23	(0.75, 1.89)	1.25
12	1.48	(0.71, 2.73)	1.61
13	1.56	(1.15, 2.07)	1.53

- **Issue 3: Measurement Level Agreement**
  - **Concern: Reliability and Validity Testing Reported at the Case Level Rather than the Hospital Level**
    - Our submission included information on reliability testing at the case level. As the reviewers noted, we sought to test the reliability and validity of the predictor and outcome variables ultimately used in the model to report the metric standardized adverse event ratio.
    - In response to the review, we have tried to provide additional information at the site level. In the response to the audit (Issue #4), we have provided reliability testing for the classification of predictor and outcome variables at the Hospital level not included in our original submission.
    - We have also looked for center variability in the severity classification of major adverse events detailed in the response to Issue #5.
    - In addition, in response to Issue #8 we provide institutional data and testing of the sample institutions' representativeness of all groups.
  - **Concern: There was No Validity Testing of the Metric at the Hospital Level to Another of the Same Construct**
    - The metric for site comparison of adverse outcomes was developed because there is no current gold standard to assess the quality of institutions that perform congenital cardiac catheterization. We could not identify another metric at the institution level against which to test the performance of the CHARM II metric. Some have proposed surrogates, such as case volume, but these surrogates do not adequately account for variation in case mix or patient complexity, as the proposed metric is designed to address.
- **Issue 4: Response to Questions Related to the Audit and the Cohort Sample**
  - The audit sample of 650 cases was ***randomly*** selected from the entire cohort and not limited to cases with an adverse event. Case ascertainment and database recording was verified by

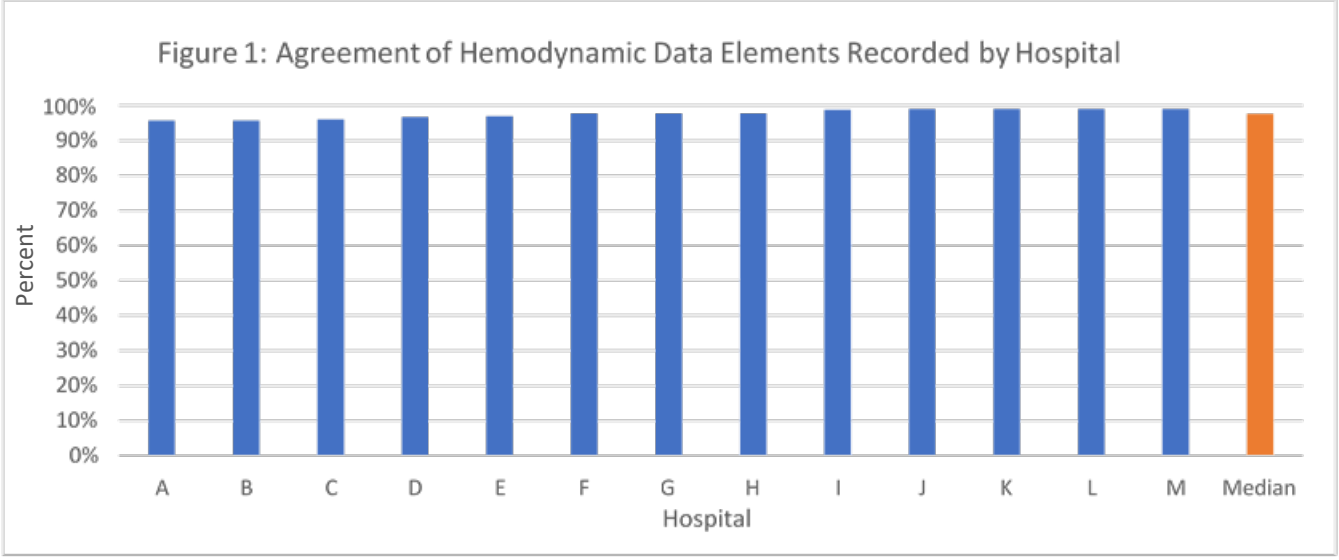
matching case volume to institutional records for total cases with 97 to 99% agreement among the institutions in the cohort.
- Reporting reliability was addressed by auditing a random sample of cases at each site. Although the sample for the audit is small relative to the total sample size (3%), it did include review of 650 randomly selected cases from the cohort.
- In addition to auditing the occurrence of adverse events, patient and procedural characteristics in the model were audited and are now included in the response to Issue #6.
- Reporting reliability is detailed in the submission for both major adverse events (the outcome for the metric; severity level 4,5) and clinically significant adverse events (severity level 3,4,5). In summary, among the 650 audited cases, 26 of 27 clinically significant level 3, 4, and 5 adverse events were recorded in the database. Thus, the audit suggests minimal recording bias across all hospitals in reporting the occurrence of major adverse events.
- A summary of the comparison of the audit sample to the cohort was not previously provided and is now included in **Appendix 2.**
- 2018 and 2019 data were not included in the analysis as the audit of AE severity classification was performed at the end of 2018 and the metric components were built in the calendar year of 2019 and submitted in its current form in January 2020 to NQF.

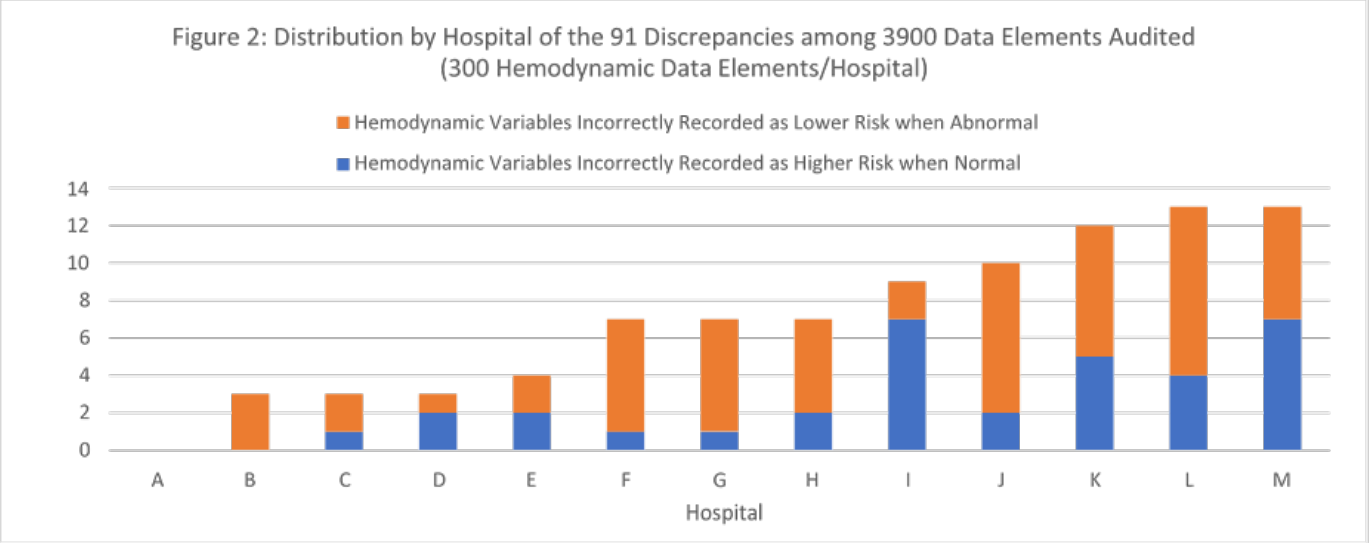
♦ **Issue 5: Outcome *Adverse Events does not* include *Testing of Chance Corrected Agreement***

- All adverse events were independently reviewed by two interventional cardiologists for proper classification of severity level categorization.
- In the 2014-2017 data set, 3094 adverse events of all severity levels were recorded. Among 267 adverse events classified as severity level 4 by the site at which the event occurred, 11 (4.1%) were downgraded to level 3 upon review. Among 962 adverse events classified as severity level 3 by the sites, 73 (7.6%) were upgraded to level 4; among 1635 events classified as level 2, 10 (0.6%) were upgraded to level 4.
- In total, 94/3094 (3.0%) adverse events were recategorized in a manner which would affect the outcome "any major (level 4/5) adverse event." At the case level, 9 cases were reclassified from having a major adverse event to not having one, and 83 cases were reclassified from not having major adverse event to having one. Prior to the audit, the rate of any major adverse event would have been reported as 1.1% instead of the current 1.4% in the cohort.
- Note that we are unable to report a kappa statistic to assess reliability of the outcome of any major adverse event. All cases for which the sites did not report any adverse event at all were not audited for the purpose of identifying events. If, however, we assume that all cases for which an adverse event was not reported by the site would have been determined not to have a major AE upon review, the kappa would be 0.85

♦ **Issue 6: Information on Reliability Testing for Risk Factors Not Provided**

- The predictor variables were also assessed in the audit but were not reported in our submission. Please allow us to provide these important results for your review:
  - For ***procedure type*** and ***age***, there was 100% agreement in the audited data set across centers.
  - For the ***hemodynamic indicator variables***, among 3900 variables audited, 57 were recorded incorrectly in a lower risk category and 34 were recorded incorrectly in a higher risk category compared to the audit results. Thus, there was 97% (3809/3900) agreement in reported versus source document audited data. By center the distribution of reliability in recording ranged from 96 to 99% (**Figure 1**). **Figure 2** below shows the number of incorrectly recorded hemodynamic variables as higher risk when normal and lower risk when abnormal by center.





Validity

Validity Assessment of Threats to Validity

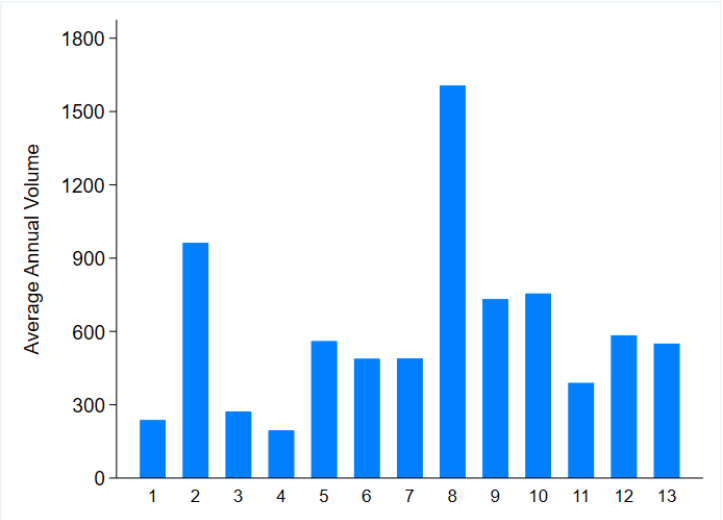
- **Issue 7: Risk Adjustment**
  - **Concern: Lack of Justification for using Procedure Type Risk Group in the Model**
    - Because many different types of procedures are performed in congenital cardiac catheterization, we created procedure type risk categories over a decade ago to group commonly and uncommonly performed procedures in groups with similar expected outcomes, minimizing variation within categories while maximizing variation between categories.
    - In previous work, procedure type risk category has been found to be the most significant explanatory variable for the outcome adverse events. Empirically, this is also the case for the new procedure type risk categories in this data set with a c statistic of 0.68 in univariate analysis.
    - The reviewers raise concerns that procedure type is not known prior to the procedure, however the metric is not intended to be used for prediction. Rather, the risk adjustment model is used to adjust for population level case mix complexity to allow for

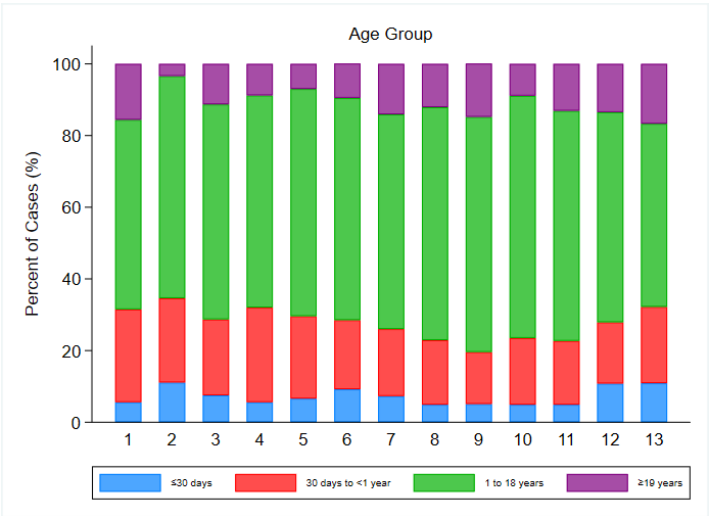
equitable retrospective comparison of major adverse outcomes across institutions.

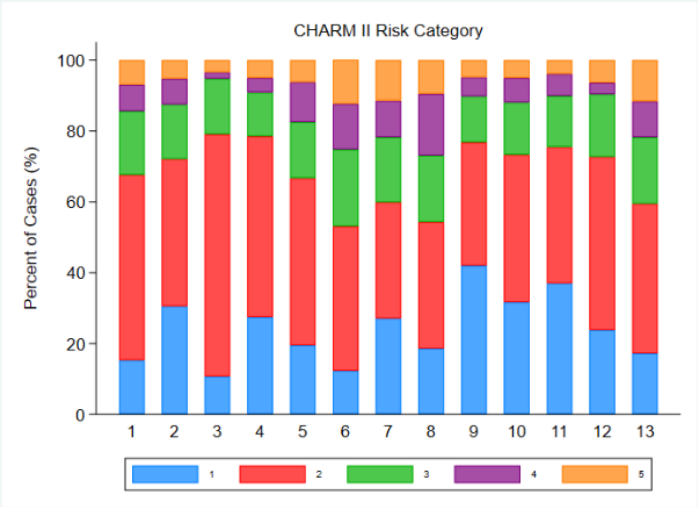
- **Concern: Lack of Social Risk Factor Adjustment**
  - We share the reviewers' dissatisfaction with the lack of adequate social risk factor assessment and adjustment in the model. However, we are limited by the data collected in this multi-center database cohort where SES data was not collected at the patient level.
  - To date, a multi-center dataset that can support such an analysis has not been pursued in congenital cardiac catheterization and, consequently, potential influencers of patient outcomes have not been explored or reported.
  - In a retrospective effort to address this gap, we explored the impact of insurance status at the institution level by proportion of government insured patients on the outcome of our model. However, the lack of insurance status data at the case level did not allow for adequate assessment of the impact socioeconomic circumstances on patient outcomes. The direction of our result was not consistent with SES analysis in other populations and meaningful conclusions cannot be drawn given the limitations of available data in this SES analytical attempt.
  - The developers acknowledge that the methodology for assessment of SES factors on patient outcomes is inadequate and that our database was not structured to provide meaningful analysis focused on social determinants of health. We hope this is a domain that can be explored in future datasets or built upon our work with a subsequent consideration of the impact when considered in the model.

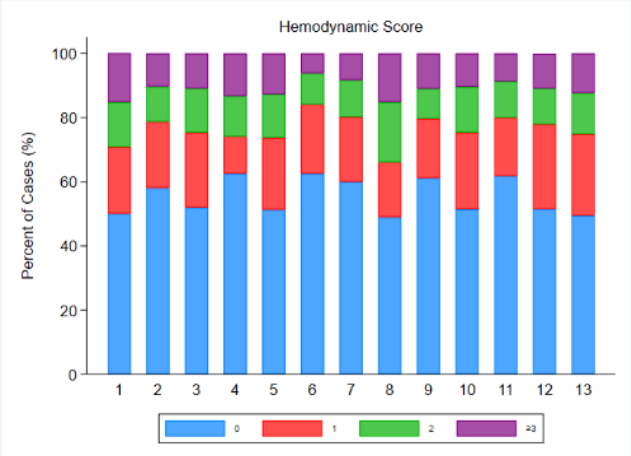
Validity Assessment: Additional Threats to Validity

- **Issue 8: Concern Regarding Hospital Representativeness of all Groups Measured**
  - Additional information is provided in this response regarding hospital volume and frequency of the model predictors by hospital, **Appendix 3** and **Figures 3-6** below.



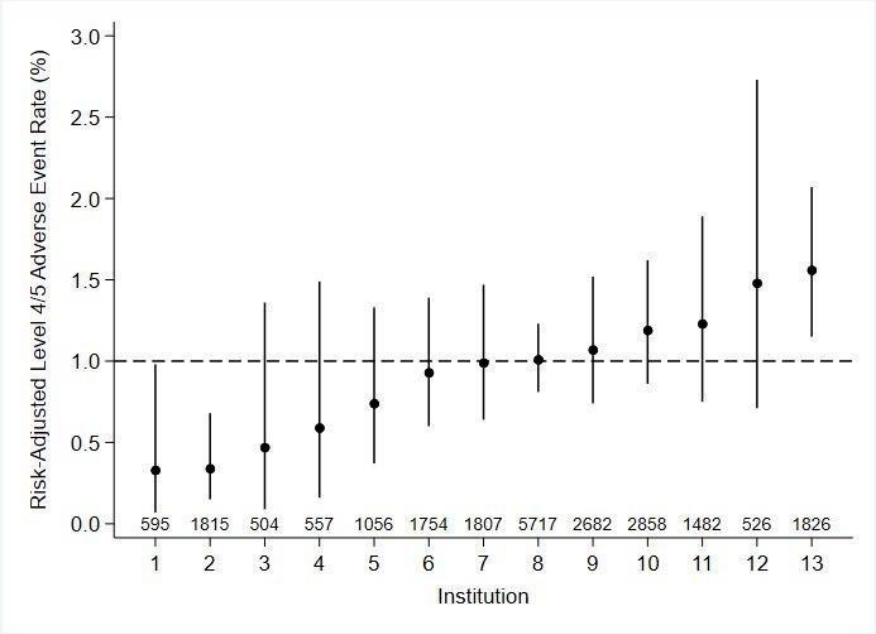






♦ **Issue 9: Metric Performance**

- **Concern: Ability to Identify Meaningful Differences in Hospital Performance**
  - Per reviewer suggestion, we have added sample sizes for each site on the SAER figure. Please see the revised SAER figure below (**Figure 7**), with hospital case volumes included. The case volumes provided show the number of eligible cases used in the SAER calculation from each hospital.
  - Figure 7: Revised SAER with Hospital Case Volume



- **Concern: The Testing Set Only Included 13 Hospitals**
  - The IMPACT (IMproving Pediatric and Adult Congenital Treatment) Registry is the largest national registry, organized by the ACC (American College of Cardiology) NCDR (National Cardiovascular Data Registry). This registry has 106 participating institutions in the United States. According to the IMPACT 2019 Quarter 3 report, 37,352 cases were recorded in the registry in four quarters, or one year. Median annual volume for participating centers is approximately 320 cases per year.
  - Thus, although the sample size is interpreted as small, the annual volume of this data set represents approximately 15-20% of all congenital heart disease catheterization cases performed annually in the US.
  - Furthermore, the cohort used in developing this metric includes representation from small, medium, and large volume centers, with annual volumes ranging from approximately 200 to 1600. Thus, based on center volume, the metric development cohort is representative and generalizable to the intended population.

♦ **Appendix 1: CHARM II Patient and Procedural Characteristics: January 2014 through December 2017**

- Comparison of procedures with case type = 0 versus those with other case types values shown are number (percent) or median [25th, 75th percentiles]

	Analyzed Cohort (n=23212)	**Case Type 0** (n=2290)	**P Value**
Age (n=23179, 2288)			<0.001
≤30 days	1552 (7)	160 (7)	
>30 days to <1 year	4442 (19)	442 (19)	
1 to 18 years	14508 (63)	1294 (57)	

≥ 19 years	2677 (12)	392 (17)	
Sex Male (n=22989, 2280)	12650 (55)	1311 (58)	0.024
Weight (kg) (n=23149, 2276)	17.0 [8.0, 47.8]	22.0 [7.8, 58.6]	<0.001
Single Ventricle (n=23152, 2008)	4614 (20)	533 (27)	<0.001
Genetic Syndrome	2157 (9)	171 (7)	0.003
Any Non-Cardiac Problem	4337 (19)	435 (19)	0.72
Cardiac Cath in Last 90 Days	4179 (18)	560 (24)	<0.001
Cardiac Surgery in Last 90 Days	3187 (14)	517 (23)	<0.001
Cardiac Intervention in Last 90 Days	5351 (23)	770 (34)	<0.001
Hemodynamic Score			<0.001
0	12707 (55)	1597 (70)	
1	4692 (20)	332 (14)	
2	3142 (14)	205 (9)	
≥3	2671 (11)	156 (7)	
Any Level 4/5 Adverse Event	321 (1.4)	45 (2.0)	0.034

- **Appendix 2: CHARM II Patient and Procedural Characteristics: January 2014 through December 2017**
  - Comparison of audited procedures to those that were not audited Values shown are number (percent) or median [25th, 75th percentiles]

	Not Audited (n=22562)	**Audited Cases** (n=650)
Age (n=22530, 649)		
≤30 days	1504 (7)	48 (7)
>30 days to <1 year	4295 (19)	147 (23)
1 to 18 years	14133 (63)	3751 (58)
≥ 19 years	2598 (12)	79 (12)
Sex Male (n=22345, 644)	12296 (55)	354 (55)
Weight (kg) (n=22500, 649)	17.0 [8.0, 47.9]	15.8 [7.4, 45.3]
Single Ventricle (n=22503, 649)	4479 (20)	135 (21)
Genetic Syndrome	2079 (9)	78 (12)
Any Non-Cardiac Problem	4181 (19)	156 (24)
Cardiac Cath in Last 90 Days	4071 (18)	108 (17)
Cardiac Surgery in Last 90 Days	3090 (14)	97 (15)
Cardiac Intervention in Last 90 Days	5200 (23)	151 (23)
CHARM II Risk Category		
1	5688 (25)	161 (25)

2	8991 (40)	278 (43)
3	3800 (17)	113 (17)
4	2327 (10)	51 (8)
5	1738 (8)	47 (7)
Hemodynamic Score		
0	12284 (55)	363 (56)
1	4555 (20)	137 (21)
2	3072 (14)	70 (11)
≥3	2592 (11)	79 (12)
Any Level 4/5 Adverse Event	314 (1.4)	7 (1.1)

- **Appendix 3: CHARM II Patient and Procedural Characteristics by Site: January 2014 through December 2017**

SITE		1	2	3	4	5	6	7	8	9	10	11	12	13
TOTAL NUMBER OF CASES		711	1924	544	584	1120	1950	1961	6425	2927	3021	1554	584	2198
AVERAGE ANNUAL VOLUME		237	962	272	195	560	488	490	1606	732	755	389	584	550
AGE	≤30 days	6%	11%	8%	6%	7%	9%	7%	5%	5%	5%	5%	11%	11%
	31 days to <1 year	26%	24%	21%	26%	23%	19%	19%	18%	14%	19%	18%	17%	21%
	1 to 18 years	53%	62%	60%	59%	63%	62%	60%	65%	66%	68%	64%	59%	51%
	≥ 19 years	16%	3%	11%	9%	7%	10%	14%	12%	15%	9%	13%	14%	17%
CHARM II RISK CATEGORY	1	15%	31%	11%	28%	20%	12%	27%	19%	42%	32%	37%	24%	17%
	2	53%	42%	69%	51%	47%	41%	33%	36%	35%	42%	39%	49%	42%
	3	18%	15%	16%	12%	16%	22%	18%	19%	13%	15%	14%	18%	19%
	4	8%	7%	2%	4%	11%	13%	10%	17%	5%	7%	6%	3%	10%
	5	7%	5%	3%	5%	6%	12%	11%	10%	5%	5%	4%	7%	12%
HEMODYNAMIC SCORE	0	50%	58%	52%	63%	51%	63%	60%	49%	61%	51%	62%	52%	49%
	1	21%	21%	23%	12%	22%	22%	20%	17%	19%	24%	18%	26%	25%
	2	14%	11%	14%	13%	14%	10%	12%	19%	9%	14%	11%	11%	13%
	≥3	15%	10%	11%	13%	13%	6%	8%	15%	11%	10%	9%	11%	12%
ANY LEVEL 4/5 ADVERSE EVENT		0.5%	0.4%	0.6%	0.7%	1.0%	1.4%	1.4%	1.6%	1.2%	1.4%	1.4%	1.9%	2.6%

Measure #3576 Pediatric Asthma Emergency Department Use (UCSF)

Summary:

Thank you for the detailed and careful review of our testing documentation.

There were a few key weaknesses that we believe we have addressed in our response. These include issues with reliability testing, low R-squared values, and the need for validation and calibration testing in a separate dataset from the dataset in which the measures were originally developed.

We have addressed each of these issues, the first two by changing the level of analyses. Our dataset for measure calculation is structured as a member-month dataset (each line represents one month of the measurement year for one member, with up to 12 lines per member, since each month requires eligibility assessment, and each month has a variable for number of ED visits for that member for that month). We performed our original testing to calculate ICC and R-squared calculation using the member-month outcome (how many ED visits for a member in any given month). As noted, this resulted in low ICC and low R-squared, since asthma-related visits are relatively rare and hard to predict for individual children, and even more so for any given month. In considering reviewers comments and suggestions, we agree with the need for plan-level calculations of ICC and R-squared, rather than using the member-month assessment. These are the results presented in the following responses to specific comments.

We respond to the request for validation and calibration testing in a separate dataset by conducting the full set of analyses in CA data. These results are also presented below in response to individual comments.

There is also some confusion regarding our validity testing in the VT dataset. We apologize for not being clearer in describing the outcome measure used for the analysis. We present the data to support the face validity and usability evidence that this measure can be used to inform quality improvement efforts. For health plans interested in improving the measure, a learning collaborative of practices is a very concrete method to improve performance. So though the analysis is not at the plan level, it provides additional strength to the application in illustrating that improvement on the measure is associated with improvements in asthma care processes that are under the direct control of clinicians.

In this document, we have also responded to additional specific comments from reviewers.

Thank you for your review and guidance.

Reliability

Specifications

- **Issue 1:** I do not understand why qualifying events differ by age. A content expert may be able to expound on this. Do all ED visits/admissions count of just one per child per observation time assessed.
    - **Developer Response 1:** These age differences are per NHLBI guidelines (https://www.nhlbi.nih.gov/health-topics/guidelines-for-diagnosis-management-of-asthma) and were reviewed and developed in collaboration with the Delphi panel of experts convened during the development of this measure. Due to the short timeline on this response to reviews, and the ongoing competing demands due to the COVID19 pandemic, we are not able to elaborate in greater detail.
- **Issue 2: Complexity of measure specifications.** I find the explanation of the calculation of the numerator in 100 child years (S.5-S.7 and S.14) to be very confusing (which, in itself, means I do not think they will be reliably calculated by health plans). I think that in addition to the text examples provided, there should also be example calculations to ensure all understand.

- **Developer Response 2:** We can provide example calculations in the measure submission. We also have SAS code that is publicly available with documentation, in order to assist health plans in implementing these measures using existing data.

Testing

- **Issue 3: Multiple reviewers commented on the low ICC (“**The intraclass correlation coefficient test is poor as noted in the table provided on p.6 of the testing form.” “The “MA health plans” ICC was 0.00076 and “CA health plans” ICC was 0.0029. Typically, we want to see a result above 0.5 in the ICC test to consider the result reasonably reliable.” **“**ICCs were quite low (ranging from 0.00076 to 0.0039) reflecting a serious reliability problem, as one might expect given the rarity of asthma ED admissions. However, the results in section 2b4 show that the measure is effective at detecting meaningful differences, which does not seem to square with the results of the ICC analysis” and “Estimated ICC as the proportion of the variation in outcome is due to the group (i.e. health plan) being evaluated. This is not a measure of reliability. Reliability can either be estimated using (1) SNR or (2) split-sample reliability testing. In split-sample reliability testing, (1) the sample is randomly split into two halves, (2) the performance of each group is estimated in each of the two data samples, an (3) the two sets of measures are then compared using the ICC. This “ICC” is different from the one estimated by the measure developers.”
  - **Developer Response 3**: We agree with reviewer comments noting that estimating individual risk of an ED visit on any given month for an individual presents reliability problems due to the rarity of the events. We also appreciate the suggestion that it is more appropriate to analyze plan level ICCs rather than patient-level ICCs, in recognition that this is a plan-level measure and therefore the need is to demonstrate plan-level reliability, rather than patient level reliability.
  - In response, we have re-run the data using the suggested split sample approach, using both CA and MA data (we were not able to perform these analyses with VT data in the short time frame, as we do not have direct access to VT data).
  - Our updated table for Health Plan reliability using the same set of variables and statistical approach to risk-adjustment, assessing health plan performance reliability using a split sample approach as suggested.

Table 1. Reliability ICC testing using split sample analysis and ICC calculation of plan performance for split samples.

Level of testing	ICC	Confidence interval	Number of clusters	Number of patients	Number of patient-months*
MA Health Plans	0.72	0.49-0.86	26 plans	83,577	698,420
CA Health plans	0.86	0.79-0.90	101 plans	321,072	3,098,769

*We use patient-month here for consistency within the table, but this is the same as member-month.

- These results show that when assessing reliability at the plan level, that the measure has moderate to high reliability. This is reassuring to us and provides evidence of the strength of the measure reliability. This also is consistent with our findings that the

measure is effective at detecting meaningful differences, which was noted by the reviewer above to be previously inconsistent. This further suggests that plan-level reliability calculation is the correct approach.

- **Issue 4:** In addition, in 2a2.4, the developer indicates "It is possible for us to also assess plan and clinic-level ICCs in VT APCD data, in light of the improvement in performance at the clinic level (noted below in Section 2b1). We have not done that analysis but plan to do so and so could make those results available upon request," and I think that information would provide additional information that could be helpful to assess reliability.
    - **Developer Response 4:** In light of the results above, and due to short time line and not having direct access to the VT data, we did not run the analysis in the VT APCD data.

Validity

Measure exclusions

- **Issue 5:** No concerns apart from the fact that exclusions were tested using only data from MA. Are these results representative of other states?
    - **Developer Response 5:** In response to this comment, we assessed missingness in CA data. Results are as follows:
    - Data was complete for age, sex, and chronic condition indicator for all patients.
    - Data on social risk factors was missing for 0.53%-0.58% of patients.
    - The level of missingness differed across plans with a high of 3.31% and a low of 0.
    - Due to the low level of missingness, we did not conduct further sensitivity analyses. Our interpretation of this analysis is that the level of missingness in CA is not substantial.

Ability to identify meaningful differences in performance.

- **Issue 6:** I could not understand which sample was used for the performance analysis. The N seems to be 29. Is this a sub-sample?
    - **Developer Response 6:** This is the sample of plans. See Table 1 above in Issue 3.
- **Issue 7:** The data reported are on practice differences in response to a practice-level asthma QI improvement intervention, not a plan-level analysis (Table 2). It is unclear from data for 2b4.2 how the plan categories (n-29) were classified into high, no difference from average, low performing groups were identified, nor the assertion that "40% of plans identified as high or low performing" establishes clinically meaningful differences.
    - **Developer Response 7: Our description of how we determined outlier status is in 2b4.1, and the results are reported in 2b4.2 and interpreted in 2b4.3, as requested in the testing document. We used the following methods to identify outliers:**
    - We used standard z-score methodology to identify high, medium and low performers, based on the CMS approach to identifying high, medium and low performers in their public reporting programs. In order to use the z-score methodology of identifying outliers, we did the following: We first fit a mixed effects negative binomial regression model with random effects for payer and fixed effects as noted above. We then generated the predicted effects and standard errors for each plan, in a post-estimation command. We then calculated the Z-statistic for each plan, using the predicted effect and standard error for each plan.

- Plans with a Z-statistic>1.96 were considered poor performing outliers, those with <-1.96 were considered high performing outliers and those in between were considered no different from average.

- **Issue 8:** Reliability testing at the plan level was done on CA and MA samples but not VT. In contrast, validity testing was done on VT sample (albeit taking advantage of a practice-level QI intervention). No plan-level validation appear to have been done.
  - **Developer Response 8:** We have now performed a plan level validation analysis of the risk-adjustment model developed in MA data, using the CA data. See Issue XX.

Missing data

- **Issue 9:** No concerns apart from, again, not being clear on which sample was used for testing impact of missing data.
  - **Developer Response 9:** We used the MA dataset for missingness analysis in the testing attachment. See above for new CA missingness analysis results in Issue 5.
- **Issue 10:** Plan with 40% missing data was dropped. Social risk data missing on 7%. Three plans with 100% missing data. unclear how this will be managed. AND: Concerned with the degree of missing SES data, especially given its inclusion as a risk variable. It was missing for 6.6% of cases. Of the 26 plans sampled (on p. 20), half (13) of the plans had 10% or more missing SES data.
  - **Developer Response 10:** We apologize for the confusion. Our testing attachment submission response on missing data presented numbers at the member-month level. Below we present a table with data at the member/patient level. Based on this data, we suggest dropping plans from measurement without social risk factor data available for at least 40% of observations. In MA, this would only exclude 371 patients (0.43% of patient sample). In addition, standard public reporting methods do not recommend including entities with less than 25 eligible patients,

Table. Distribution of missing data across plans

Plan ID	Percent observations with missing social risk factor data	Total patients in plan
11715	0%	15
12226	0%	12
7397	0%	3
290	0%	3
3156	1%	17,776
3735	1%	13,447
3505	1%	7,762

4962	1%	6,670
11541	1%	178
296	2%	2,162
301	3%	4,930
8026	6%	564
10632	7%	1,377
300	9%	10,870
8647	10%	4,220
302	11%	942
7041	13%	45
10440	18%	53
312	20%	503
10441	20%	335
291	24%	13,257
11474	26%	196
10353	47%	40
10444	50%	228
11939	100%	85
11943	100%	9
11936	100%	9

- **Issue 11:** I apologize but I do not fully understand the statistical approach to determine outlier status for plans.
    - **Developer Response 11:** As noted above in Issue 7, we used standard z-score methodology to identify high, medium and low performers, based on the CMS approach to identifying high, medium and low performers in their public reporting programs. In order to use the z-score methodology of identifying outliers, we did the following: We first fit a mixed effects negative binomial regression model with random effects for payer and fixed effects as noted above. We then generated the predicted effects and

standard errors for each plan, in a post-estimation command. We then calculated the Z-statistic for each plan, using the predicted effect and standard error for each plan.
    - Plans with a Z-statistic>1.96 were considered poor performing outliers, those with <-1.96 were considered high performing outliers and those in between were considered no different from average.
- **Issue 12:** It is not clear from the analyses presented whether assessment of missing data was performed for the measure. Data presented are for social risk factors. It is also not clear which sample(s) the data represent (CA, MA, both, VT?)
    - **Developer Response 12:** The assessment of missingness was done for all variables included in the measure. Data was complete for all other variables in the model: age, sex, and chronic condition indicator for all patients. ED visit data was available for all patients, though we are not able to assess whether any ED visit claims were missing for patients, which is a similar limitation to other NQF endorsed measures (e.g., readmissions).

Risk adjustment

- **Issue 14:** Unknown whether "medical comorbidity status" is the person's disposition at the "start of care", or whether the status is assigned based on the measurement year. Given this is a health plan measure, I'm assuming "start of care" (as noted in the question) means the beginning of the measurement year.
    - **Developer Response 14:** Correct.
- **Issue 15:** The sample used for the risk-adjustment development is almost 700,000 patients, noted to be data from MA. This does not correspond to the MA data shown in section 1.6. Please clarify.
    - **Developer Response 15:** We apologize for the confusion. Both numbers are correct. The number of observations reporting in the STATA output reflects the member-month observations in the dataset. The data in section 1.6 shows the number of patients included, which is the member count, not the member-month count. Note that the number of patients is roughly 1/12 the number of member-month observations, though not precisely since not every patient contributed all 12 months of data.
- **Issue 16:** The low predictive power of the risk-adjustment model (R-sq near zero) question its utility. What is the justification to use a model with such low predictive power?
    - **Developer Response 16:** As noted above, the original analyses (R-squared and ICC) were calculated using the member-month file. We have re-run the R-squared analysis on the plan-level performance assessment, to assess the predictive power for that model. To conduct this analysis, we performed a linear regression used a plan-level dataset, with the performance of the plan as the outcome, including the mean values of the variables of interest across all the member-months.
    - **New R-squared values**: 0.56 for MA data; 0.13 for CA data.
    - The interpretation is that using the MA data, the model explains 56% of the variance in the outcome (56%) and 13% of the variance in the outcome using the CA dataset (13%).
- **Issue 17:** From what I can tell, no c-statistic was reported but the R squared is very, very low (0.0023) suggesting this model is not effective.

  - **Developer Response 17:** The c-statistic is not appropriate since the measure is a rate, not a proportion. See above Issue 16 for repeated analysis assessing plan level R-squared with a higher number suggesting that the model explains 13-56% of variance, suggesting that the model is more effective for plan-level assessments.
- **Issue 18:** I like that the developer gathered additional social risk factors that were not available in the claims data. These included: 1) % households below the poverty level; 2) % population with less than high school education; and 3) % male unemployment for 25 to 60 year olds. However, I did not see a discussion of why race/ethnicity was not included as I would hypothesize that this would also be a strong predictor of socioeconomic disadvantage. Also the low R2 concerns me a bit. What other co-variates should be in the model to improve the predictability?
  - **Developer Response 18:** Since race/ethnicity are not direct measures of social risk, it is not recommended per NQF and ASPE guidance to use them. See Issue 16 for our response to the low R2 (R-squared).
- **Issue 19:** Several reviewers were looking for cross-validation or risk-adjustment assessment in another dataset to assess representativeness, and one reviewer was confused regarding the additional sensitivity testing we presented in the optional testing for risk adjustment section.
  - **Developer Response 19:** We have repeated all testing, including risk adjustment, ICC, R-squared, calibration, and identification of outliers in the CA dataset. See below in the Validity testing section for these results.

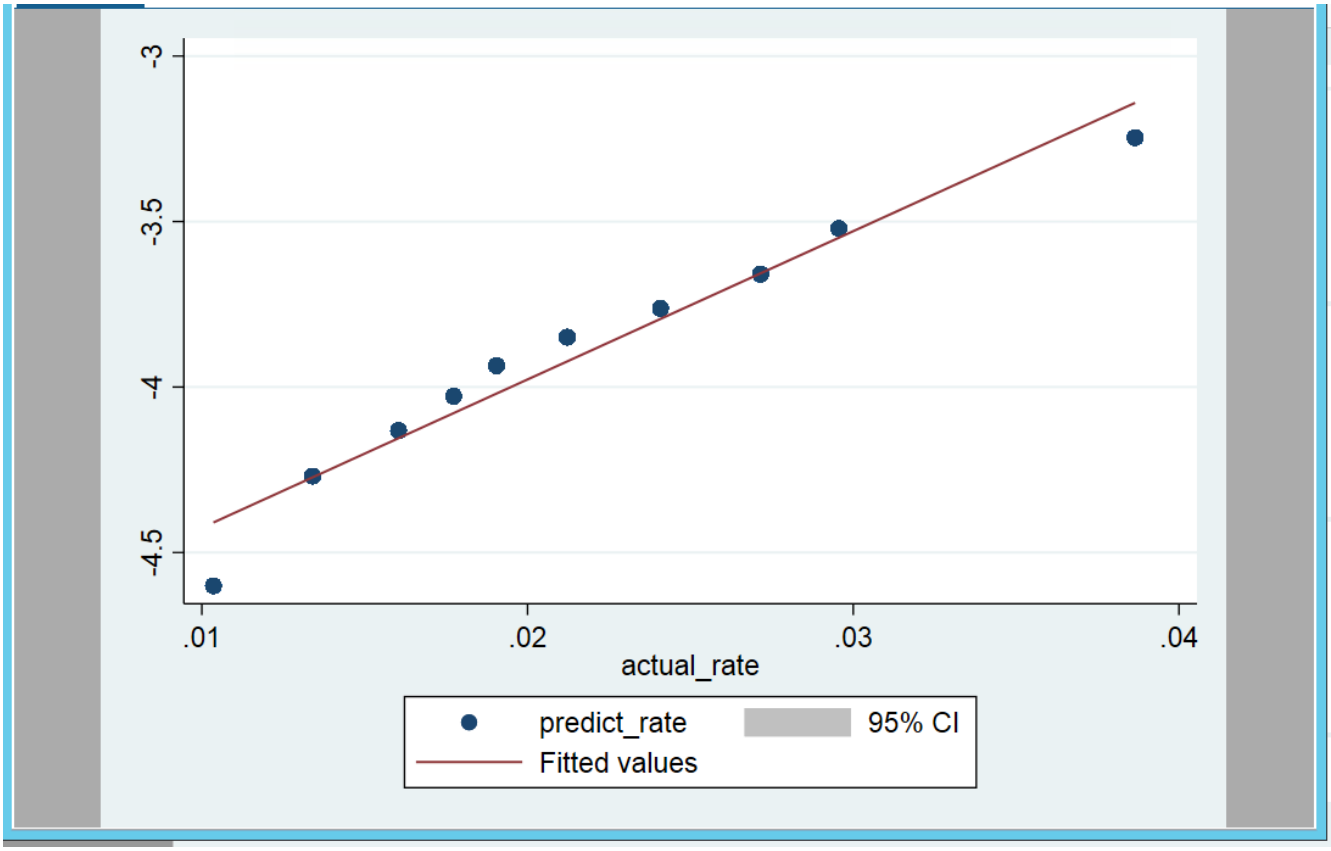
Validity testing

There were two major concerns regarding the data we presented in validity testing. An issue that arose from multiple reviewers was the lack of validation in an additional dataset and the validation of the measure in VT data using clinic data only. We respond to this concern by rerunning the data in CA and present the results below.

The other concern was that the validation data we present was done at the clinic level, not the plan level, and that it does not support the measure itself, as it did not use the measure specifications. We respond to this with a reframing of the VT analysis as evidence of face validity of the measure (the measure can be improved through a QI collaborative of the type a health plan might organize in order to improve asthma-related ED visit rates). The empirical validity testing we present using the CA data now, not the VT data.

- **Issue 20:** The model appears to have excellent calibration on the calibration curve. However, it does not appear that the model was validated in a validation data set. Model validation was performed using the same data used to develop the model. This is a major limitation.
  - **Developer Response 20:** In order to test the model in a validation dataset, we calculated ICC, R-squared, calibration, and outlier analysis using the CA dataset, thus testing the model in a different dataset from the one in which it was developed. ICC is presented in our response to Issue 3 and R-squared in our response to Issue 16. The calibration and outlier results for CA are below. They both use the same approaches as described in the testing document section 2b3.5. for calibration, and in Issue 7 above for outlier analysis**,** but with CA data:

- Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*): **R=0.98** (correlation coefficient for predicted vs actual rates by decile of predicted)
- Statistical Risk Model Calibration – Risk decile plots or calibration curves:



- Note: the predicted and actual_rate values in the graph are not transformed. Transformed values are reported below in 2b3.9.
- 2b3.9. Results of Risk Stratification Analysis:

Decile	Predicted	Actual
1	12.1	12.4
2	16.8	16.1
3	19.3	19.2
4	21.4	21.3
5	23.5	22.9
6	25.6	25.5
7	27.9	28.9
8	30.9	32.6

9	35.5	35.5
10	46.7	46.4

- Results of outlier analysis using CA data:

outlier	Freq.	Percent
High Performing	16	15.38
No different from Average	56	53.85
Low performing	32	30.77
Total	104	100.00

- **Issue 21:** b) In addition, the model does not appear to include an offset term for the exposure (=the number of patient-year asthma per group).
  - **Developer Response 21:** The use of an offset term would be appropriate if the data structure was data at the payer level, with the outcome being the number of admissions for the payer for the month, in which case one would want to know the denominator- the number of patient-years of asthma for the payer (since 72 admissions out of 72 patients in a month is way different from 72 admissions out of 7200 patients). However, because the data structure for performance calculation is in the person-month level, we do not need the offset term for the modeling.
- **Issue 22:** The validity testing looked at if the measure was responsive to QI initiatives implemented in clinics, but the measure is specified for health plans, not clinics. Validity testing based on a health plan-level measure would be stronger. Validation assessment appears to have been done at the practice not plan-level.
  - **Developer Response 22:** See our response to Issue 20 above.
- **Issue 23:** Disagree with the validity testing employed where they assessed whether there was a change in ED use of children with asthma pre and post a QI effort in Vermont. In 2b1.2, it did not necessarily state the measure being discuss here was necessarily employed to measure ED use in this study. Thus, this study design does not allow us to conclude anything specifically about the "Pediatric Asthma Emergency Department Use" (#3576) measure. AND: The developers used data from a Vermont QI learning collaborative focused on improving asthma care and management in primary care to assess whether this intervention helped decrease

asthma-related ED utilization. This is an indirect indication of the measure's validity in the aggregate, but it says nothing about the measure's ability to distinguish among plans.

- **Developer Response 23:** We apologize for not being clearer in describing the outcome measure used for the analysis. The data presented used technical specifications of the proposed measure #3576, though slightly modified. It was an earlier analysis and so did not include social risk factors or the comorbidity variable in the risk adjustment model. However, it followed the rest of the technical specifications, including the same numerator and denominator definitions, rate definition, and ICD codes and rolling look-back period for identifiable asthma as well as insurance eligibility requirement. We present the data to support the face validity and usability evidence that this measure can be used to inform quality improvement efforts. For health plans interested in improving the measure, a learning collaborative of practices is a very concrete method to improve performance. So, though the analysis is not at the plan level, it provides additional strength to the application in illustrating that improvement on the measure is associated with improvements in asthma care processes that are under the direct control of clinicians.

- **Issue 24:** The sample used is not clearly described. What is the 'N' for the validity testing? Does table 2 in section 2b1.3 describe an analyses done on a patient or a practice level?
  - **Developer Response 24:** There were 20 practices participating in the learning collaborative and 15 control practices. See Table below for patient demographics included in the analysis. The analyses used the member-month dataset. We used the following approach: To assess the relationship between asthma QI collaborative participation and asthma-related ED utilization, we used a difference-in-differences approach. This approach compared the asthma-related ED visit rate per 100 child-years before (2014) and after (2017) the QI collaborative at participating versus control practices. We used a mixed-effects negative binomial multivariable regression model, accounting for clustering within patient and practice. To estimate the difference-in-differences effect, models included variables for participation, year, and the interaction term between participation and year. The p-value for the interaction term tested whether the change in ED utilization was different between participating and control practices. We obtained adjusted ED rates at each time point by using the post-estimation margins command.

Table: Characteristics of patients at participating and control practices before (2014) and after (2017) the asthma quality improvement learning collaborative

	2014					2017				
	Participating		Control			Participating		Control		
	N	%	N	%	*P*-value	N	%	N	%	*P*-value
Total Study Sample	2376	100	1282	100		2257	100	899	100	
Age Categories										

3 to 5 years	453	19	206	16	0.047	430	19	160	18	<0.001
6 to 11 years	935	39	498	39		868	38	326	36	
12 to 17 years	839	35	478	37		854	38	327	36	
18 to 21 years	149	6	100	8		105	5	86	10	
Gender										
Male	1193	50	644	50	0.26	1125	50	442	49	0.32
Female	1046	44	580	45		1009	45	419	47	
Unknown	137	6	58	5		123	5	38	4	
Insurance										
Non-Medicaid	897	38	412	32	0.001	627	28	183	20	<0.001
Medicaid	1479	62	870	68		1630	72	716	80	

Notes: CHAMP: Child Health Advances Measured in Practice; N=sample size; %=percentage; p-values are from bivariate Chi-squared tests comparing patient-level percentages within demographic categories across participant and control practices within each year.

- **Issue 25:** Although not specified as such, the measure seems to have good face validity. Some elaboration on this under a face validity section would be helpful to further support the measure's overall validity.
    - **Developer Response 25:** This is beyond the scope of this response but we agree about the face validity of the measure, based on the VT data presented above as well as extensive asthma literature. We can elaborate in the full measure submission.
- **Issue 26:** Validity testing was conducted using only the Vermont data, excluding data from CA & MA. No description was provided for the VT data in section 1.6. No sensitivity analyses were done to demonstrate the representativeness of the VT data compared to CA & MA.
    - **Developer Response 26:** We have now presented validity data from running the measure in CA data, as well as presenting demographic data from the VT dataset in the Table above.
- **Issue 27:** As noted above, there is no direct assessment of validity of the measure scores compared to scores from a similar construct.
    - **Developer Response 27:** This was not a specific request or requirement in the testing attachment items, and is beyond the scope of this response. However, we would like to note that the intention in presenting the VT analysis is to support the face validity and usability of this measure, demonstrating that it can be used to inform quality improvement efforts. For health plans interested in improving the measure, a learning collaborative of practices is a very concrete approach to improve performance. The

analysis provides additional strength to the measure application by illustrating that improvement on the measure is associated with improvements in asthma care processes that are under the direct control of clinicians.

- **Issue 28:** Above results support hypothesis that QI initiative lowers the rate captured by this measure. Do the developers have measure reliability data on this particular set of observations? That would help set my mind at ease regarding low reliability
    - **Developer Response 28:** See our response above in Issue 3 regarding low reliability.

Appropriate method

- **Issue 29:** I was torn between moderate and low, based more on reliability than validity testing. Also tempted to suggest insufficient because the reliability of the score from the VT study would have helped
    - **Developer Response 29:** Please see repeat reliability testing above in Issue 3.
- **Issue 30:** The model appears to have excellent calibration on the calibration curve. However, it does not appear that the model was validated in a validation data set. Model validation was performed using the same data used to develop the model. This is a major limitation.
    - **Developer Response 30:** We have repeated the calibration analyses in a separate dataset. See Issue 20.
- **Issue 31:** Examined impact of QI intervention using a difference-in-difference model using negative binomial regression. This approach is not adequate for demonstrating the validity of the measure itself. It can show that the QI intervention has an impact on the outcome of interest. This analysis does show that the outcome is responsive to the QI intervention. But it, by itself, cannot be used to validate the measure itself. In particular, the fact that this intervention led to improvement in outcomes does not ensure that this measure appropriately adjusts for case mix – and should be used to publicly measure health plan performance.
    - **Developer Response 31:** We have now conducted validity testing in a separate dataset (CA Medicaid data) from the dataset in which it was developed (MA APCD data). See Issue 20.

Measure #2687 Hospital Visits after Hospital Outpatient Surgery (Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE))

Reliability

- **Issue 1: Measure Specifications:** Methods Panel reviewers asked clarifying questions about the measure's specifications, including:
    - *How or whether urgent care visits within 7 days are counted.*
    - *Does every overnight care episode with AM discharge count in the numerator?*
    - **Developer Response 1:**
        - Urgent care visits within 7 days are not counted in this measure.
        - The outcome, hospital visits, include inpatient, observation stays, and emergency department visits. So yes, if billed as an observation stay or inpatient stay, every overnight care episode with AM discharge would count, unless it is a potentially planned admission, or within one of the other exclusion categories

(such as same day/same claim or same claim after an emergency department [ED] visit).

- **Issue 2: Planned Readmissions:** A Methods Panel reviewer requested information about *how the algorithm determination of planned/unplanned fit in actual circumstances?*
    - **Developer Response 2:** We are interpreting this question to be asking about how accurate the algorithm is in practice. The algorithm is widely used and working well in practice. CMS and CORE initially formally validated the algorithm (see attached manuscript).[k] In addition, the algorithm is currently used across CMS' hospital readmission measures and outpatient measures of hospital visits post procedures, and revised in response to user feedback. Its patient-level results are routinely shared with providers as part of CMS public and confidential reporting; all hospitals, for example, get patient-level data for each CMS readmission measure that indicate whether each patient had an unplanned readmission (as determined by the algorithm). CMS also maintains Q&A in-boxes for measures, and addresses any feedback on the algorithm through annual reevaluation. Finally, CMS updates the algorithm annually to incorporate changes to procedure and diagnosis codes.
- **Issue 3: Gaming and unintended consequences.** A Methods Panel reviewer asked two questions:
    - *Does the existing CMS list prevent a surgeon from listing every procedure as inpatient (for admission) and then switching to same day discharge after the fact to avoid registering a numerator/denominator event?*
    - *"Pain" is a very common reason for seeking care after surgery (including POD0, especially for obese patient with chronic pain). Is there additional documentation about unintended consequences of such a measure, particularly in the midst an opioid epidemic?*
    - **Developer Response 3:**
        - We do not expect surgeons to keep patients out of the measure through designating patients as inpatients initially. CMS audits billing records for inpatient surgeries with patient stays that are below the two-midnight benchmark.[l] Therefore if a provider designated patients as inpatients initially and then discharged them on the same day, that would be below the two-midnight benchmark, the claim would be flagged for CMS review, and the claim could potentially be denied if none of the exceptions applied.
        - While pain is one of the reasons patients return to the hospital for care after a procedure, it is not the most common reason. For example, for several of the procedures in this measure (for example, those in the body system groups of urinary, skin/breast, respiratory, male genitalia, and others) pain is not among the top 10 reasons patients return to the hospital (see Attachment A). Given the current national and clinical focus on proper opioid use we do not believe that this measure will incentivize excess opioid prescribing; however, we appreciate the committee's flagging this concern, and CMS will consider monitoring for this potential unintended consequence during measure reevaluation. CMS also regularly surveys providers about its quality programs, in part regarding barriers

[k] Horwitz LI, Grady JN, Cohen DB, et al. Development and validation of an Algorithm to Identify Planned Readmissions From Claims Data. J Hosp Med. 2015;10(10):670–677.

[l] https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNMattersArticles/downloads/MM10080.pdf. Accessed March 13, 2020.

and unintended consequences of implementing its quality measures.[m3] As this measure was first publicly reported in January 2020, CMS does not yet have feedback from providers on unintended consequences of this specific measure.

- **Issue 4: Risk adjustment variables.** One Methods Panel reviewer stated that the measure's risk model variables were not visible in the measure information form (MIF).
    - **Developer Response 4:** CORE submitted the risk model variables in Table 1 on pages 19 and 20 of the testing attachment. We provide them again in the table below.

Table 1: Logistic Regression Model Variable Odds Ratios (January 1, 2018-December 31, 2018; Dataset #2)

Parameter	Odds Ratio	95% CI
Age minus 65 (years above 65)	1.02	1.02-1.03
Comorbidities:		
Cancer (CC 8-14)	1.02	1.00-1.03
Diabetes and DM Complications (CC 17-19, 122, 123)	1.15	1.13-1.17
Disorders of Fluid/Electrolyte/Acid-Base (CC 24)	1.15	1.13-1.17
Intestinal Obstruction/Perforation (CC 33)	1.17	1.13-1.17
Inflammatory Bowel Disease (CC 35)	1.07	1.00-1.13
Bone/Joint/Muscle Infections/Necrosis (CC 39)	1.37	1.32-1.44
Hematological Disorders Including Coagulation Defects and Iron Deficiency (CC 46, 48, 49)	1.12	1.10-1.14
Dementia or Senility (CC 51-53)	1.18	1.15-1.21
Psychiatric Disorders (CC 57-63)	1.15	1.13-1.17
Hemiplegia, Paraplegia, Paralysis, Functional Disability (CC 70, 71, 73, 74, 103-105, 189-190)	1.18	1.14-1.22
Other Significant CNS Disease (CC 77-80)	1.18	1.14-1.22
Cardiorespiratory Arrest, Failure and Respiratory Dependence (CC 82-84)	1.06	1.03-1.09
Congestive Heart Failure (CC 85)	1.13	1.10-1.15
Ischemic Heart Disease (CC 86-89)	1.14	1.12-1.16
Hypertension and Hypertensive Disorders (CC 94, 95)	1.08	1.06-1.10
Arrhythmias (CC 96, 97)	1.13	1.11-1.15
Vascular Disease (CC 106-109)	1.14	1.12-1.16
Chronic Lung Disease (CC 111-113)	1.13	1.11-1.15
UTI and Other Urinary Tract Disorders (CC 144, 145)	1.14	1.12-1.15
Pelvic Inflammatory Disease and Other Specified Female Genital Disorders (CC 147)	0.90	0.86-0.93
Chronic Ulcers (CC 157-161)	1.10	1.06-1.13
Cellulitis, Local Skin Infection (CC 164)	1.17	1.14-1.19
Prior Significant Fracture (CC 169-171)	1.41	1.37-1.45

[m] https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/National-Impact-Assessment-of-the-Centers-for-Medicare-and-Medicaid-Services-CMS-Quality-Measures-Reports. Accessed March 13, 2020.

Morbid Obesity (CC 22)	1.15	1.12-1.19
Work Relative Value Units	1.12	1.11-1.12
Body System Operated On:		
Cardiovascular	1.99	1.77-2.23
Digestive	3.34	2.98-3.74
Ear	Reference	
Endocrine	1.86	1.64-2.12
Female Genitalia	2.85	2.51-3.24

Parameter	Odds Ratio	95% CI
Hemic-Lymphatic	2.10	1.78-2.48
Skin & Breast	1.47	1.31-1.65
Male Genitalia	3.75	3.34-4.21
Miscellaneous Procedures	1.03	0.42-2.52
Musculoskeletal	2.39	2.13-2.68
Nervous	2.99	2.67-3.36
Nose-Throat-Pharynx	2.59	2.27-2.95
Respiratory	2.56	1.99-3.29
Urinary	3.89	3.47-4.36

- **Issue 5: Facility volume threshold for public reporting:** One Methods Panel reviewer was concerned that CMS would lower the facility volume threshold for public reporting, which is currently set at 30.
    - **Developer Response 5:** If changes in the measure score reliability suggested the need to change the volume threshold, CMS would need to announce this change through formal rulemaking, which includes a public comment period for stakeholders to provide CMS with feedback on the proposed change.

Validity

- **Issue 1: Exclusions:** One Methods Panel reviewer expressed concern about the exclusion of surgeries that were on the same day and same claim as the surgery (about 5% of the cohort). The reviewer stated that *"It would be useful to see what would be the impact on measure score if these are included with assumption that they are all in the numerator."*
    - **Developer Response 1:** Due to the short turn-around time for CORE's response to these questions, we are unable to provide the Methods Panel with this result. However, it may be possible to run this analysis and have it prepared for the Standing Committee review of the measure.
- **Issue 2: C-statistic:** One Methods Panel reviewer expressed concern that the c- statistic for this measure was below 0.7.
    - **Developer Response 2:** We suggest the committee interpret the c-statistic in the context of this particular measure. If an outcome is more strongly related to quality of care rather than patient characteristics, patient factors are less predictive of the outcome. The results from our variable selection suggest that for this measure, patient history has a relatively limited relationship to the occurrence of a hospital visit within 7 days; as supported by the conceptual model for the measure and the literature, the outcome is also predicted by other factors, such as the quality of care delivered by the

facility. Note that the c-statistic for this measure, 0.684, is higher than other similar recently-NQF- endorsed measures.

- **Issue 3: External empiric validity:** One Methods Panel reviewer provided advice on additional empiric validation analyses. **"***Other forms of validity testing as known groups validity could have been used to better assess the measure score validity. For example, hospitals with a case mix of patients with higher rates of characteristic expected to be associated with better outcomes would be expected to have higher scores compared to hospitals with higher rates of case-mix of patients with characteristic expected to be associated with worse outcomes.*"
    - **Developer Response 3:** We thank the Methods Panel reviewer for this suggestion. To restate the suggestion, the reviewer suggests we compare measures scores for hospitals (let's call them Group A) that have case-mixes of patients with patient characteristics associated better outcomes (patients with fewer risk factors/comorbidities) with hospitals that have case mixes of patients with patient characteristics associated with worse outcomes (Group B). The reviewer states that we would expect better performance from Group A hospitals compared with Group B hospitals.
      This analysis would not serve to validate the measure, however, given that the measure is risk-adjusted for case mix. The numerator is the number of admissions predicted by the hierarchical model among a hospital's patients, given the patients' risk factors and that hospital's hospital-specific effect. The denominator is the expected number of admissions among that hospital's patients given the patients' risk factors and the average of all hospital-specific effects in the nation. CMS takes the ratio of the numerator and denominator explained above (predicted/expected or P/E), to calculate the risk-standardized hospital visit ratio (RSHVR).
      We risk adjust so expected values reflect differences in case mix. The hospital's expected values will be higher if they have a higher-risk case mix, while they will be lower if they have a lower-risk case mix. The hospital will be worse or better than expected based on the hospital's contribution to the outcome (the hospital- specific quality effect, in the numerator). Therefore, you could have a Group B hospital with a higher predicted than a Group A hospital, but because the hospital has a worse case-mix, its expected value will also be higher, and therefore it could have the same performance (measure score) as a Group A hospital with "healthier" (lower-risk) patients, and a lower expected value. In summary, risk- adjustment allows hospitals with varying case mixes to perform better or worse than what would be expected with their particular case-mix.
- **Issue 4: Empiric external validity:** One Methods Panel reviewer stated that they *"Disagree with the premise that the measure being evaluated would be necessarily correlated with the HWR measure. Rationale is twofold: [1] These are differing units of analysis, [2] The measure steward states "It is possible the same surgeons and surgical teams are performing surgeries covered by both measures". However, no analysis/ evidence of this is noted as to the degree to which this occurs. Additionally, it is not only the surgeon that influences the outcomes, but numerous other factors as well (e.g. the team, systems in place in each setting)."*
    - **Developer Response 4:** We thank the Methods Panel reviewer for their input and agree with the comment. On page 11 of the testing attachment, we hypothesize the measure scores would be weakly positively correlated. We note that [1] it is possible that the same surgeons and surgical teams are performing surgeries covered by both measures, and in some hospitals those procedures may be co-located, [2] both measures count admissions to the hospital post- surgery in the outcome, although the HOPD measure also counts ED visits, which make up the majority of the return visits, as well as observation stays, and [3] the same organizational culture and processes may be in place to prevent visits to the hospital following surgery across both inpatient and outpatient procedures, such as timely recognition of post-operative complications and ensuring

effective discharge plans. However, as the reviewer notes, we did not provide an analysis of overlap of the surgeons or surgical teams. While we are currently unable to directly analyze this overlap, there is evidence from the literature that individual surgeons perform both inpatient and outpatient surgeries,[n] and we can consider such an analysis in future reevaluation.

- **Issue 5: External empiric validity.** Methods panel reviewers expressed concern about the external empiric validity of the measure.
    - **Developer Response 5:** As further support for empiric validity, we submit, for review by the Scientific Methods Panel, the top 10 reasons for return visits to the hospital following outpatient surgery, stratified by body system group (see Attachment A). Most of the reasons for return listed for each body group are related to the surgery, including complications of surgery. For example, the top 10 reasons for revisit for the digestive group, accounting for 37% of hospital visits, include urinary retention (9.9%), constipation (2.4%), acute post-operative pain (2.3%), and surgical complications such as hemorrhage (2.4%), hematoma (2.0%), and other surgical complications (7%). In contrast, our TEP decided to not include eye surgeries in part because the reasons for return were not considered to be related to the surgery.
- **Issue 6: Face validity.** One methods panel reviewer stated that *"It would have been more helpful if the measure steward would have provided us with other survey results."*
    - **Developer Response 6:** We have provided all of the results that are available.

Measure #2496 Standardized Readmission Ratio (SRR) for dialysis facilities (University of Michigan Kidney Epidemiology and Cost Center)

Reliability

- **Issue 1: There were several concerns and questions raised about the interpretation and application of the IUR and PIUR and particularly about the differences in IUR between this current submission and the previous 2014 submission based on 2009-2012 data.**
    - **Developer Response 1:** We acknowledge the IUR declined from the previous 2014 submission (0.34 current versus 0.55 previous submission). There are multiple reasons for this change, including changes in the underlying data source, and the impact of the implementation of the measure in CMS public reporting and value-based purchasing programs. We outline these below.
        1. Since the last submission, the SRR was implemented as a quality measure in both the ESRD QIP value-based purchasing program, and the public reporting Dialysis Facility Compare (DFC) and DFC Star Ratings. These programs have incented greater quality improvement attention from facilities. Therefore, one might expect to see a move overall to adopt better practices, which would potentially reduce the between facility variation and therefore reduce the IUR.
        2. In 2012 Medicare claims reporting of comorbidities expanded from 10 to 25 condition diagnoses. This allowed for greater reporting and broader universe of comorbidity risk adjustment that in turn may have reduced the between provider variation, resulting in lower IUR.
        3. Since the 2014 submission of SRR, ICD coding transitioned from ICD-9 to the more granular ICD-10 diagnostic codes. Additionally, we shifted the identification of

[n] Darrith B, Frisch NB, Tetreault MW, Fice MP, Culvern CN, Della Valle CJ. Inpatient Versus Outpatient Arthroplasty: A Single-Surgeon, Matched Cohort Analysis of 90-Day Complications. J Arthroplasty. 2019;34(2):221–227.

patient prevalent comorbidities from the CMS Hierarchical Condition Categories (HCC) to the clinically derived AHRQ CCS diagnosis categories.

4. For the current submission, we restrict to only inpatient Medicare claims to assess comorbidities. The original SRR in 2014 ascertained comorbidities from both inpatient and outpatient claims. Restricting to inpatient claims allowed us to reduce bias by using inpatient claims for Medicare Advantage patients, since outpatient Medicare claims are not available for the Medicare Advantage subpopulation. This transition to restricting to inpatient claims also harmonizes our approach to comorbidity assessment with the NQF #1789 All Cause Hospital Wide Readmission measure implemented in CMS programs. This ICD coding transition could also decrease variation between facilities. For example, if we do not include comorbidity adjustments in the model, the IUR increases from 0.34 to 0.44.
5. We have improved our method of estimating the effects of comorbidities. To estimate the effects of comorbidities, we use a logistic regression model with facility-hospital combinations included as fixed effects while adjusting for patient-level characteristics. The estimates of the regression coefficients from this model avoid issues of bias that arise when estimates are obtained using a model with random effects where biased estimates occur when the hospital or facility effect is correlated with the covariates.
6. The original 2014 reported IUR included all readmissions within 30 days. In response to stakeholder comments we modified the SRR to exclude readmissions within 0-3 days since discharge. For this reason as well, the original and current IUR are not directly comparable.

Each of these changes likely impacts the value of the IUR. We also posit that our re-evaluation work for SRR yielded a substantially better model that is able to account for more of the between facility differences that would in turn reduce the IUR (i.e., signal to noise between facilities).

- **Issue 2: Questions about interpreting the IUR and the PIUR.**
  - **Developer Response 2:** Kalbfleisch et al. (2018) explains that the interpretation of the IUR as reliability depends on the differences between providers being entirely (or mostly) due to the quality of care. In many if not most instances, however, this is not the case. There are differences in the patients treated by providers that are not accounted for in the adjustments that we are able to make. In effect, there will almost always be unmeasured confounders that are related to the outcome and also differ between facilities. For example, these include genetic differences among the patients treated that vary across facilities, or dietary differences, or differences in the level of family support, etc. We do not measure these variables, but they are undoubtedly important and they contribute to the between facility variation. Thus, one can have a high value of IUR due simply to incomplete risk adjustment and in general, adjusting for confounders can reduce the IUR. Similarly, an IUR near 0 does not mean that the measure is not useful for profiling. In fact, if most of the providers have outcomes centered very near a national average while a relatively smaller number have outcomes that are out of line, the IUR would be near 0, and yet the measure may be very useful for identifying those extreme facilities. For these reasons, the IUR should be interpreted with care as it may not reflect the true reliability of the measure.
    These considerations motivated the definition of the PIUR, which concentrates on the ability of the measure to consistently flag the same facilities. The PIUR is introduced in He et al. (2019), where additional examples can be found. We have also used this measure in our

submissions for the standardized mortality ratio (SMR) and the standardized hospitalization ratio (SHR). In particular, the overall IUR for the four-year SMR (2015-2018) is 0.50. The corresponding PIUR of the four-year SMR is 0.77. The overall IUR for 2018 SHR is 0.53. The corresponding PIUR of 2018 SHR is 0.75. In many instances, one is particularly interested in identifying providers whose outcomes are extreme and the PIUR concentrates on this aspect.

Note that the PIUR is very close in spirit to the definition of reliability in the testing form*: "**2a2. Reliability testing** demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise."*

- **Issue 3: Several questions were raised about the bootstrap methodology for computing the IUR.**
  - **Developer Response 3:** In nonlinear models such as the binary logistic model used here, the usual ANOVA formulas cannot be used to obtain the IUR. For this reason, we developed a method (He et al., 2017) that uses a bootstrap approach to estimate the within provider variation of a given measure for use in the IUR computation. The IUR itself is still a measure of the proportion of the variation that is accounted for by differences among providers. This is a more efficient approach than the data splitting approach that has been used by some developers, and relatively simple to implement. We also understand that these methods have now been adopted by at least one other developer.

    There was also a question as to why we have not presented the IUR separately for different facility sizes. We made this change since we had noted that most other developers did not do this and when we did present this separation, discussion often centered on the lowest value (for the smallest facilities). The breakdown, however, by tertiles of total patient-years are as follows:

2015 SRR

Patient-years	**Overall**	**[0,39.3]**	**(39.3,66.4]**	**(66.4,314]**
IUR	0.3667919	0.2773175	0.3693323	0.4189114
PIUR	0.6118983	0.5439296	0.6402836	0.6399477

2016 SRR

Patient-years	**Overall**	**[0,38.8]**	**(38.8,65.8]**	**(65.8,314]**
IUR	0.3499017	0.2277934	0.3388834	0.4358279
PIUR	0.5647885	0.3453441	0.5055602	0.7149113

2017 SRR

Patient-years	Overall	[0,37.6]	(37.6,64.8]	(64.8,314]
IUR	0.3634977	0.245504	0.3469823	0.4468958
PIUR	0.6688346	0.4871348	0.6573758	0.7784522

2018 SRR

Patient-years	Overall	[0,35.8]	(35.8,63.6]	(63.6,314]
IUR	0.3453658	0.2734061	0.305310	0.4187057
PIUR	0.6077232	0.5482808	0.5315561	0.7049255

The higher PIUR compared to the IUR (overall and across all groups with various sample sizes) indicates the presence of providers with extreme outcomes, a feature that is not captured in the IUR itself. Also, unlike the IUR, the PIUR is not necessarily larger for facility groups with larger sample size. Even a group of facilities with a very low IUR can have relatively high PIUR and the corresponding measure can be very useful for identifying providers in this group with extreme outcomes.

References:

1. Kalbfleisch JD, He K, Xia L, Li Y. Does the inter-unit reliability (IUR) measure reliability?, Health Services and Outcomes Research Methodology, 2018 Sept. 18(3), 215-225. Doi: 10.1007/s10742-018-0185-4.
2. He K, Dahlerus C, Xia L, Li Y, Kalbfleisch JD. The profile inter-unit reliability. Biometrics. 2019 Oct 23. doi: 10.1111/biom.13167. [Epub ahead of print]
3. He K, Kalbfleisch JD, Yang Y, Fei Z. Inter-unit reliability for nonlinear models. Statistics in Medicine. 2019 Feb 28;38(5):844-854. doi: 10.1002/sim.8005. Epub 2018 Oct 18.

Validity

- **Issue 1: A few panel members expressed concern about the coefficient estimates indicating strength of association between SRR and other primary and intermediate outcomes: standardized hospitalization ratio (SHR), standardized mortality ratio, long-term catheter (vascular access), standardized fistula rate (SFR, vascular access).**
    - **Developer response 1:** While the Pearson correlation coefficients were lower than in the prior submission we emphasize that the hypothesized associations (correlation coefficients) are in the expected direction and all highly significant at the $p<0.0001$ level ($p<0.0006$ for SRR and LTC). We do not consider the declines to be substantial, particularly given the many changes in the underlying data and each of the measure definitions.
      Data: The testing for the original SRR submission in 2014 used 2009 data which pre-dates the transition to the ICD-10 diagnoses codes (used for prior year comorbidity risk adjustment in SRR). The validity testing correlations included in the current submission uses data after the transition to ICD-10 for SRR and the measures used for correlation analysis

(i.e. SMR, SHR, STrR, SFR). Additionally, the prior year comorbidities in the 2014 submission of SRR were based on the HCC groupers, while the current SRR uses the clinically derived AHRQ CCS groups.

Measure changes: The 2019 SHR, SMR, and vascular access measures used in the empirical validity testing for the 2020 reevaluation were notably different with respect to risk adjustments and population being measured. The SHR and SMR used in the 2014 submission only adjusted for comorbidities at ESRD incidence while the current production version of these measures used in the 2020 testing include adjustment for 210 prevalent comorbidities; SMR in 2014 was an all–patient measures versus the current SMR which is restricted to the Medicare population. Both the vascular access measures used in the 2014 testing (unadjusted fistula rate, unadjusted catheter > 90 days rate) were claims based measures and restricted to the Medicare population, while the current LTC and SFR are CROWNWeb based measures and include all patients; SFR is also adjusted for a set of prevalent and incident comorbidities; and both SFR and LTC included exclusions for limited life expectancy. Finally, the SHR, SMR, LTC, and SFR used in the current testing with SRR are the 2019 production versions (as calculated and released on Dialysis Facility Compare) and do not yet reflect our updated method for handling of Medicare Advantage patients that was applied to the current SRR under review.

In light of these changes, it is perhaps not surprising that there are relatively larger changes in the correlation coefficients observed. We maintain, however, that the expected and consistent direction of the hypothesized relationships, general magnitude of the coefficients, and the statistical significance of the associations with SRR demonstrate stability from the previous to the current empirical validation results. Therefore we argue that the empirical validity testing results are both stable and robust to changes and updates since 2014, which in our assessment provides validation support for SRR and its empirical association with other primary and intermediate outcomes.

- **Issue 2: Inclusion of Medicare Advantage patients in the measure calculation (note: issue also raised under Reliability).**
    - **Developer Response 2:** Our review of the Methodology Panel's evaluation of the current SRR submission suggests that several members of the Methodology Panel review team have reservations about the measure developer's decisions related to continued inclusion of Medicare Advantage (MA) patients in the measure calculation. We believe some of these reservations may have been rooted in uncertainty about our methodology for inclusion of claims-based comorbidities for both MA and fee-for-service (FFS) Medicare patients. We sincerely apologize for any ambiguity in the submission that may have contributed to the panel's uncertainty. We hope that this response clarifies our methodology for reducing the potential bias that can result from the presence of two very different Medicare payment models in the target population.

    In the current SRR Testing Form, we included the following justification: *The SRR is dependent on Medicare claims and other CMS administrative data for several important components of measure calculation, including identification of comorbid conditions. For these reasons, the SRR was originally developed and, subsequently implemented as, a measure limited to Medicare patients.*

    For several Medicare-only measures developed by UM-KECC, the presence of active Medicare coverage has been defined using a combination of criteria including a defined minimum of paid claims for dialysis services and/or presence of a Medicare inpatient claim during an eligibility period. With the recent increase in Medicare Advantage (MA) coverage

for Medicare chronic dialysis patients, and the known systemic issue of unavailable outpatient claims data for MA patients, these criteria have the potential to introduce significant bias into measure calculations that could affect results for dialysis facilities with either very low or high MA patient populations.
As part of the comprehensive measure review process, we assessed the extent of MA coverage for ESRD dialysis patients and the effect of our historical definition of "active Medicare" status on the measure result. Medicare Advantage patient status was defined using Medicare Enrollment Database (EDB) criteria. Primary Medicare Fee for Service (FFS) coverage was identified using CMS administrative data, and active Medicare status utilized the combination of minimum dialysis paid claims and/or inpatient Medicare hospitalization claims briefly described above. We confirmed the presence of usable ICD diagnosis codes from MA inpatient claims and the nearly complete absence of outpatient Medicare claims data for patients identified as MA in the CMS data used for our measure calculation.
Summary findings:

1. *The percentage of patients with MA coverage receiving chronic dialysis in US dialysis facilities has approximately doubled in the last decade and is approaching 20% based on 2017 data.*
2. *We confirmed the presence of usable ICD diagnosis codes from MA inpatient claims and the nearly complete absence of outpatient Medicare claims data for patients identified as MA in the CMS data used for our measure calculation*

Additional analyses (Table 5) demonstrate a variable distribution of Medicare Advantage ESRD dialysis patient proportion following geographic boundaries. For example, the percentage of MA ESRD patient time at risk relative to total Medicare ESRD patient time at risk varies from a low of 2.2% in Wyoming to a high of 44.2% in Puerto Rico.

Based on the above results, we have included Medicare Advantage patients in the measure, but have limited the identification of comorbidities to inpatient claims (which are available for patients of all insurance types) and added an adjustment factor to account for Medicare advantage patients in the model. This minimizes risk of biased results at the dialysis facility level and is consistent with a number of other NQF-endorsed measures that are based on Medicare claims data.

In this response we add the following points to clarify our decision-making regarding submitted measure changes:

1. Medicare Advantage patients have been included in the SRR since CMS began requiring Medicare Advantage providers to submit non-payment claims for acute hospitalization care under the Medicare Inpatient Claims system. Specifically, beginning January 7, 2008, hospitals were required to begin submitting "no pay" claims to their Medicare contractor for stays by Medicare Advantage beneficiaries in order for CMS to calculate Disproportionate Share Hospital calculations for eligible hospitals (MLN Matters 5647, http://www.cms.gov/MLNMattersArticles/downloads/MM5647.pdf September 2012). The presence of inpatient claims data (non-paid) for MA patient hospitalizations have never been excluded from the index hospitalization identification process. Therefore we are not adding MA patients to the revised SRR since they were already included through our identification of eligible inpatient claims. Rather, we are eliminating an emerging source of bias directly caused by the rapid growth of Medicare Advantage insurance coverage in the US dialysis population and our measure denominator, and accounting for the known absence of outpatient and Physician Supplier claims data for MA patients.

2. For the purposes of identifying co-morbidities from Medicare Claims, the exact same process is used for both Medicare FFS and MA patients. We utilize all available inpatient claims from the index discharge AND from other inpatient claims in the 12 months prior to the index discharge for both FFS and MA patients. We no longer use outpatient claims sources to identify co-morbidities, in an attempt to eliminate potential bias related to the near universal absence of outpatient claims for MA patients. As one would expect, identification of prevalent comorbidities based on only inpatient claims results in fewer comorbidities for each patient compared to use of the universe of Medicare claims. However, use of only inpatient claims results in similar numbers and types of comorbidities for MA patients and other Medicare patients. For instance, in an analysis of a set of comorbidity groups used in a recent SRR calculation, we found that inpatient claims identified 12 comorbid conditions for MA patient on average compared to 12.4 comorbid conditions for other (non-MA) Medicare patients.
3. One panel member raised the concern that the submitted version of SRR defines baseline comorbidities on "discharge claim only (due to lack of available prior year claims data for Medicare Advantage enrollees). The discharge claim typically includes only a subset of relevant dx and would not reflect a comprehensive risk profile (i.e., patient may have other documented dx that are not recorded on discharge claim that would increase risk level and affect expected readmissions)". As stated above, we use all available inpatient claims from the index discharge AND from other inpatient claims in the 12 months prior to the index discharge for both FFS and MA patients. While we agree that limiting co-morbidity ascertainment to inpatient claims only does, in fact, result in a less comprehensive set of co-morbidities, our recommended methodology does protect against potential bias in determining comorbidity burden due to differences in FFS and MA claim availability discussed above. Use of the inpatient claim from the index hospitalization, supplemented by other inpatient claims in the prior 365 days to define co-morbidity captures recent, likely active, comorbidities that are probably more relevant to risk-adjustment for a measure that attempts to assess care coordination in a relatively short observation window post-hospitalization. It is not certain that outpatient claim derived co-morbidities are as clinically relevant to the risk-adjustment needed for this particular measure. In addition, our approach does not require us to exclude MA patients with index discharges from the measure. We are very reluctant to eliminate 1/5 of the current observations for SRR, particularly given the anticipated growth of MA patients in the ESRD program that will result from planned changes to the MA program regulations related to ability of prevalent ESRD patients to choose MA plans beginning in 2021.
4. Our decision to use inpatient claims only for the co-morbidity risk-adjustment harmonizes with other readmission metrics with active endorsement by NQF, including the All-Cause Hospital Readmission measure (NQF#1789).

- **Issue 3: Questions about the identification of statistically significant and meaningful differences.**
  - **Developer response 3:** The estimation of SRR is based on a mixed-effects logistic regression model with fixed effects for facilities and random effects for hospitals. To test the null hypothesis that the SRR for a given facility is statistically different from the national average, we proceed in two steps.

    In the first step, we use a simulation method to calculate the nominal (one-tailed) p-value as the probability that the observed number of readmissions should be at least as extreme

as that observed given the particular patient mix in the facility and under the assumption that this facility has readmission rates corresponding to a national norm. The national norm is taken to be the median facility. Methods are described in detail in He et al. (2013) where the approach is based on the model with random hospital effects.
Flagging of facilities is based on the empirical null (Efron 2004, 2007; Kalbfleisch and Wolfe, 2013). Accordingly, the p-value for each facility is converted to a Z-score and stratified into four groups based on patient-years within each facility. The empirical null corresponds to a normal distribution fitted to the center of each Z-score histogram using a robust method. This method aims to separate underlying intrinsic variation in facility effects from variation that might be attributed to poor (or excellent) care.
Without empirical null methods, a larger number of facilities will be flagged, especially among the largest facilities. Using this method, facilities are flagged if they have outcomes that are extreme when compared to the variation in outcomes for other facilities of a similar size. The table below compares the results with and without empirical null distribution. In this case, the change in flagging rates is modest, but in other cases, the difference can be much greater.

	flagging without empirical null			
flagging with empirical null	Better than Expected	As Expected	Worse than Expected	Total
Better than Expected	136 (1.96%)	0 (0.0%)	0 (0.0%)	136 (1.96%)
As Expected	136 (1.96%)	6,216 (89.6%)	204 (2.94%)	6,556 (94.50%)
Worse than Expected	0 (0.0%)	0 (0.0%)	245 (3.54%)	245 (3.54%)
Total	272 (3.92%)	6,216 (89.6%)	449 (6.48%)	6,937

References:

1. He K, Kalbfleisch JD, Li Y and Li YJ.(2013). Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. Lifetime Data Analysis, 19(4), 490-512.
2. Kalbfleisch JD and Wolfe RA.(2013). On monitoring outcomes of medical providers. Statistics in Biosciences, 5(2), 286–302.
3. Efron B.(2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association, 99(465), 96–104.
4. Efron B.(2007). Size, power and false discovery rates. The Annals of Statistics, 35(4), 1351–1377.

Other General Comments

[Describe any additional information or considerations (that may not be related to reliability or validity) you would like the SMP to be aware of as they reconsider your measure]

- There was a comment made about stratifying on patient years rather than on number of discharges. We wanted to note that if we stratify on number of discharges, then we find that the group with the largest number of discharges will tend to have a higher percentage of readmissions. This is because hospitalization rates and readmission rates are correlated. Thus, this way of stratifying is biased by outcome dependent sampling.
- There was a comment made by one panel members about outdated references used to justify the measure; we believe they may have been reading the measure rationale section of the MIF, which was not part of the updates required for the Intent to Submit process. These references will be updated with the full submission in April.

Measure #3561 Medicare Spending Per Beneficiary – Post Acute Care Measure for Inpatient Rehabilitation Facilities (Centers for Medicare & Medicaid Services)

Reliability

- **Issue 1:** One panel member raised a concern about our use of the Minimum Data Set (MDS).
    - **Developer Response 1:** The MDS is necessary to construct one of the risk adjustment variables, indicating beneficiaries who have been institutionalized for at least 90 days in a given year. Our methodology for identifying long-term institutionalized beneficiaries is based on the CMS Part C risk adjustment model. Specifically, "*CMS uses information from the Minimum Data Set (MDS), collected routinely from nursing homes, to identify the population of long-term institutionalized. MDS assessments are sent to CMS on at least a quarterly basis. CMS uses the presence of a 90-day assessment to identify the long-term residents for payment purposes.*" CMS Medicare Managed Care Manual. Downloaded from http://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/mc86c07.pdf

- **Issue 2:** One panel member raised a concern about attribution of costs to multiple episodes and providers.
    - **Developer Response 2:** The MSPB-PAC measures are intended to encourage improved coordination of care by holding providers accountable for the Medicare resource use within an "episode of care" (episode). This episode includes the period a patient is directly under a PAC provider's care, as well as a defined period after the end of that PAC provider's treatment which may be reflective of and influenced by the services rendered by the PAC provider. Episodes, by design, may overlap with hospital and other MSPB-PAC episodes. Aligning the MSPB measures in this way is necessary to create continuous accountability and align incentives to improve care planning and coordination across all providers.

      The measures are designed to benchmark the resource use of each attributed provider against what their spending is expected to be, as predicted through risk adjustment. Consequentially, the overlap in the MSPB-PAC measures does not result in double counting.

- **Issue 3:** One panel member raised a concern about the exclusion of outliers (1st and 99th percentile) during computation of provider scores. The panel member recommended that "*the developer*

should report the distribution of outlier exclusion across facilities to ensure that they don't concentrate in a limited number of facilities."

- **Developer Response 3:** The table below provides the distribution of outlier exclusions across providers (mean, standard deviation, and 10th, 50th, and 90th percentiles). The results confirm that excluded outliers are not concentrated in a small number of providers.

  Proportion of outliers among providers

Mean	SD	P10	P50	P90
2.4%	1.8%	0.8%	2.0%	4.7%

Measure #3562 Medicare Spending Per Beneficiary – Post Acute Care Measure for Long-Term Care Hospitals (Centers for Medicare & Medicaid Services)

Reliability

- **Issue 1:** One panel member raised a concern about our use of the Minimum Data Set (MDS).
  - **Developer Response 1:** The MDS is necessary to construct one of the risk adjustment variables, indicating beneficiaries who have been institutionalized for at least 90 days in a given year. Our methodology for identifying long-term institutionalized beneficiaries is based on the CMS Part C risk adjustment model. Specifically, "*CMS uses information from the Minimum Data Set (MDS), collected routinely from nursing homes, to identify the population of long-term institutionalized. MDS assessments are sent to CMS on at least a quarterly basis. CMS uses the presence of a 90-day assessment to identify the long-term residents for payment purposes.*" CMS Medicare Managed Care Manual. Downloaded from http://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/mc86c07.pdf

- **Issue 2:** One panel member raised a concern about attribution of costs to multiple episodes and providers.
  - **Developer Response 2:** The MSPB-PAC measures are intended to encourage improved coordination of care by holding providers accountable for the Medicare resource use within an "episode of care" (episode). This episode includes the period a patient is directly under a PAC provider's care, as well as a defined period after the end of that PAC provider's treatment which may be reflective of and influenced by the services rendered by the PAC provider. Episodes, by design, may overlap with hospital and other MSPB-PAC episodes. Aligning the MSPB measures in this way is necessary to create continuous accountability and align incentives to improve care planning and coordination across all providers.

    The measures are designed to benchmark the resource use of each attributed provider against what their spending is expected to be, as predicted through risk adjustment. Consequentially, the overlap in the MSPB-PAC measures does not result in double counting.

- **Issue 3:** One panel member raised a concern about the exclusion of outliers (1st and 99th percentile) during computation of provider scores. The panel member recommended that "*the developer should report the distribution of outlier exclusion across facilities to ensure that they don't concentrate in a limited number of facilities.*"
    - **Developer Response 3:** The table below provides the distribution of outlier exclusions across providers (mean, standard deviation, and 10th, 50th, and 90th percentiles). The results confirm that excluded outliers are not concentrated in a small number of providers.

      Proportion of outliers among providers

Mean	SD	P10	P50	P90
2.3%	1.8%	0.5%	2.0%	4.2%

Validity

- **Issue 1:** One panel member raised a question about a *"reference [to] '31 days following discharge to community' rather than the 30-day post discharge presented for IRFs."*
    - **Developer Response 1:** The 31 days following discharge to community refers to the Discharge to Community (DTC) measure, not the MSPB measure, and the reference appears only in the section describing the validity analysis involving the DTC measure. The same definition of the DTC measure (with the reference to 31 days) is provided in the MSPB-PAC IRF testing attachment.
- **Issue 2:** One panel member raised a question about the risk-adjustment model for site neutral episodes. The panel member asked: "*Were the same prediction model risk factors used for both stratified models, but allowed to generate different coefficients or where different risk factors used for the two different stratifications?*"
    - **Developer Response 2:** The detailed risk adjustment specifications describe the risk factors for the Site Neutral and the Standard episodes models. Both models use similar but not identical risk factors, as described in the specifications. Specifically, the Standard episodes model does not control for Prior PAC – Institutional, Prior PAC – HHA, or Community clinical case mix categories (because standard episodes are defined as having a prior IP stay). Each model is estimating separately, with factors generating different coefficients in each model.

Measure #3563 Medicare Spending Per Beneficiary – Post Acute Care Measure for Skilled Nursing Facilities (Centers for Medicare & Medicaid Services)

Reliability

- **Issue 1:** One panel member raised a concern about our use of the Minimum Data Set (MDS).
    - **Developer Response 1:** The MDS is necessary to construct one of the risk adjustment variables, indicating beneficiaries who have been institutionalized for at least 90 days in a given year. Our methodology for identifying long-term institutionalized beneficiaries is based on the CMS Part C risk adjustment model. Specifically, "*CMS uses information*

from the Minimum Data Set (MDS), collected routinely from nursing homes, to identify the population of long-term institutionalized. MDS assessments are sent to CMS on at least a quarterly basis. CMS uses the presence of a 90-day assessment to identify the long-term residents for payment purposes." CMS Medicare Managed Care Manual. Downloaded from http://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/mc86c07.pdf

- **Issue 2:** One panel member raised a concern about attribution of costs to multiple episodes and providers.
    - **Developer Response 2:** The MSPB-PAC measures are intended to encourage improved coordination of care by holding providers accountable for the Medicare resource use within an "episode of care" (episode). This episode includes the period a patient is directly under a PAC provider's care, as well as a defined period after the end of that PAC provider's treatment which may be reflective of and influenced by the services rendered by the PAC provider. Episodes, by design, may overlap with hospital and other MSPB-PAC episodes. Aligning the MSPB measures in this way is necessary to create continuous accountability and align incentives to improve care planning and coordination across all providers.

      The measures are designed to benchmark the resource use of each attributed provider against what their spending is expected to be, as predicted through risk adjustment. Consequentially, the overlap in the MSPB-PAC measures does not result in double counting.
- **Issue 3:** One panel member raised a concern about the exclusion of outliers (1st and 99th percentile) during computation of provider scores. The panel member recommended that "*the developer should report the distribution of outlier exclusion across facilities to ensure that they don't concentrate in a limited number of facilities.*"
    - **Developer Response 3:** The table below provides the distribution of outlier exclusions across providers (mean, standard deviation, and 10th, 50th, and 90th percentiles). The results confirm that excluded outliers are not concentrated in a small number of providers.

      Proportion of outliers among providers

Mean	SD	P10	P50	P90
2.3%	2.3%	0.0%	1.7%	5.1%

Measure #3564 Medicare Spending Per Beneficiary – Post Acute Care Measure for Home Health Agencies (Centers for Medicare & Medicaid Services)

Reliability

- **Issue 1:** One panel member raised a concern about our use of the Minimum Data Set (MDS).
    - **Developer Response 1:** The information provided regarding the use of the MDS is accurate and not an error. The MDS is necessary to construct one of the risk adjustment variables, indicating beneficiaries who have been institutionalized for at least 90 days in

a given year. Our methodology for identifying long-term institutionalized beneficiaries is based on the CMS Part C risk adjustment model. Specifically, "*CMS uses information from the Minimum Data Set (MDS), collected routinely from nursing homes, to identify the population of long-term institutionalized. MDS assessments are sent to CMS on at least a quarterly basis. CMS uses the presence of a 90-day assessment to identify the long-term residents for payment purposes.*" CMS Medicare Managed Care Manual. Downloaded from http://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/mc86c07.pdf

- **Issue 2:** One panel member raised a concern about attribution of costs to multiple episodes and providers.
    - **Developer Response 2:** The MSPB-PAC measures are intended to encourage improved coordination of care by holding providers accountable for the Medicare resource use within an "episode of care" (episode). This episode includes the period a patient is directly under a PAC provider's care, as well as a defined period after the end of that PAC provider's treatment which may be reflective of and influenced by the services rendered by the PAC provider. Episodes, by design, may overlap with hospital and other MSPB-PAC episodes. Aligning the MSPB measures in this way is necessary to create continuous accountability and align incentives to improve care planning and coordination across all providers.

      The measures are designed to benchmark the resource use of each attributed provider against what their spending is expected to be, as predicted through risk adjustment. Consequentially, the overlap in the MSPB-PAC measures does not result in double counting.
- **Issue 3:** One panel member raised a concern about the exclusion of outliers (1st and 99th percentile) during computation of provider scores. The panel member recommended that "*the developer should report the distribution of outlier exclusion across facilities to ensure that they don't concentrate in a limited number of facilities.*"
    - **Developer Response 3:** The table below provides the distribution of outlier exclusions across providers (mean, standard deviation, and 10th, 50th, and 90th percentiles). The results confirm that excluded outliers are not concentrated in a small number of providers.

      Proportion of outliers among providers

Mean	SD	P10	P50	P90
1.8%	1.4%	0.0%	1.7%	3.5%

Validity

- **Issue 1:** One panel member raised the following question: *"The reference to the measure including the period "31 days after discharge" is consistent with the LTCH measure but not the IRF measure. Why?"*
    - **Developer Response 1:** The 31 days following discharge to community refers to the Discharge to Community (DTC) measure, not the MSPB measure, and the reference appears only in the section describing the validity analysis involving the DTC measure.

The same definition of the DTC measure (with the reference to 31 days) is provided in the MSPB-PAC IRF testing attachment.

Measure #3574 Medicare Spending Per Beneficiary (MSPB) Clinician (Centers for Medicare & Medicaid Services)

Reliability

- **Issue 1:** Panelists commented that no reliability testing results were provided for the clinician individual level.
    - **Developer Response 1:** Section 2a2.3 Table 1 and Table 3 of the testing form present the signal-to-noise and split-sample intraclass correlation coefficients, respectively. The rows where Reporting Level is labeled as TIN-NPI show the reliability results for the clinician individual level.

Validity

- **Issue 1:** One panelist raised concerns of assigning the same episode cost to separate TIN or TIN-NPI regardless of the number of E&M codes provided.
    - **Developer Response 1:** For episodes with medical MS-DRGs that are attributed using E&M codes, clinician groups are only attributed if they bill greater than or equal to 30% of the patient's E&Ms during the inpatient admission. For those meeting this requirement, it is important that all attributed groups share responsibility and are held under the same incentives to be cost efficient.

      All individual clinicians billing E&M services for a patient within a clinician group that has been attributed are attributed. This attribution rule recognizes the nature of team-based care in medicine and promotes the coordination of a patient's care within a clinician group while again holding all individuals involved to similar incentives.
- **Issue 2:** One panelist noted that the measure focuses on costs associated with the hospitalization period, but neglects costs during the 30 days post discharge period from the hospital. He or she raised concerns about the measures ability to influence clinicians and clinician groups to use more cost-effective PAC settings and asks how these PAC costs could be captured.
    - **Developer Response 2:** The measure does include costs in the 30 day post discharge period, with exceptions for services that are unlikely to be influenced by the clinician's care decisions. Costs of PAC services occurring during the 30 day post discharge period are included.
- **Issue 3:** One panel member questioned the exclusion of outliers (<1st and >99th percentile of residual values) during computation of provider scores. The panel member recommended that "*the developer should report the distribution of outlier exclusion across facilities to ensure that they don't concentrate in a limited number of facilities.*"
    - **Developer Response 3:** The table below provides the distribution of outlier episodes excluded across TIN and TIN-NPIs (mean, standard deviation, and 10th, 50th, and 90th percentiles). The results confirm that excluded outliers are not concentrated in a small number of clinicians or clinician groups.

      Proportion of Outliers Episodes Among Clinicians and Clinician Groups

Reporting Level	Mean	SD	P10	P50	P90
TIN	2.0%	2.7%	0.0%	1.3%	4.9%
TIN-NPI	2.6%	3.6%	0.0%	1.6%	6.7%

- **Issue 4:** One panelist ask for clarification on why the patient enrollment exclusion criteria is checked for the 90 days preceding the episode.
    - **Developer Response 4:** The risk factors included in the risk adjustment model are defined using Part A and Part B claims observed in the 90 days prior to the episode start date. Enrollment during this period is required to ensure that the risk profile of a patient is accurately and consistently defined.

Measure #3575 Total Per Capita Cost (TPCC) (Centers for Medicare & Medicaid Services)

Reliability

- **Issue 1:** Panelists commented that no reliability testing results were provided for the clinician individual level.
    - **Developer Response 1:** Section 2a2.3 Table 1 and Table 3 of the testing form present the signal-to-noise and split-sample intraclass correlation coefficients, respectively. The rows where Reporting Level is labeled as TIN-NPI show the reliability results for the clinician individual level.
- **Issue 2:** One panelist expected that the mean risk- and specialty-adjusted monthly spending by risk decile in Table 9 should monotonically increase by risk score decile.
    - **Developer Response 2:** The values in the table are risk-adjusted. As beneficiary risk increases, an increase in observed spending is expected. The risk scores however of these patients also increases in conjunction. To obtain risk-adjusted spending, observed spending is divided by the normalized risk score of a beneficiary as predicted by the CMS HCC V21 and V22 models. Risk adjustment aims to adjust the observed spending of beneficiaries relative to the expected increase resource use based on beneficiaries' present health conditions so that they can be compared. Table 9 is provided to help panelist assess the effectiveness of the risk adjustment model in accomplishing this.

Validity

- **Issue 1:** One panelist cited results from Table 6. *Share of Primary Care E&M Billed by Attributed TIN and TIN-NPI*, noting the mean share of E&M codes billed by attributed TIN or TIN-NPIs is 52.8 percent and 45.0 percent, respectively. This panelist raised concerns that the complement percentage of a patient's E&Ms (approx. 47% and 55%) are billed by non-attributed TINs and TIN-NPI. The comment also asks if this suggests that the evaluation of performance of the attributed TIN and TIN-NPIs may be based only on (approximately) half of the beneficiaries that they are attributed to.
    - **Developer Response 1:** Table 6 is a distribution across TIN/TIN-NPI *and* attributed beneficiaries. For each instance that an individual beneficiary is attributed to a

respective TIN/TIN-NPI, we calculate the proportion of all the beneficiary's E&M in the following year billed by that attributed TIN/TIN-NPI. Then, the distribution is taken across all these cases. The table best answers the question; conditional on a beneficiary being attributed to a particular TIN/TIN-NPI, how much of that beneficiary's primary care E&M codes are billed by the attributed TIN/TIN-NPI (i.e. how much E&M is a TIN/TIN-NPI billing for their attributed beneficiaries). Those claims not billed by the specific TIN/TIN-NPI are not necessarily precluded from being billed by another TIN/TIN-NPI that is also attributed the beneficiary.

Because this table is conditional on beneficiaries already attributed by the measure, we cannot infer the proportion of beneficiaries not attributed.

Measure #2539 Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy (Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE)) (Consensus Not Reached)

Validity

- **Issue 1: Social risk factor adjustment.** One Methods Panel reviewer suggested that CORE provide the results of a net reclassification analysis comparing the measure scores calculated with and without social risk factors in the model.
    - **Developer Response 1:** We appreciate the suggestion from the Methods Panel reviewer. We are not able to complete such an analysis in the short timeframe we were given for this response; however we plan on completing this analysis to have on hand during the Standing Committee meeting.
- **Issue 2: Social risk factor adjustment – c-statistic.** One Methods Panel reviewer commented that the c-statistic for both HOPDs (0.687) and ASCs (0.654) were low. The reviewer also asked about the c-statistic with the inclusion of SES, sex and race. A second reviewer noted that the results should include confidence intervals.
    - **Developer Response 2:** We suggest the committee interpret the c-statistic in the context of this particular measure. If an outcome is more strongly related to quality of care rather than patient characteristics, patient factors are less predictive of the outcome. The results from our variable selection suggest that for this measure, patient history has a relatively limited relationship to the occurrence of a hospital visit within 7 days; as supported by the conceptual model for the measure and the literature, the outcome is also predicted by other factors, such as the quality of care delivered by the facility. Note that the c- statistic for this measure is higher than other similar recently-NQF-endorsed measures.

      Sex was a candidate variable that we considered during measure development. As stated in the testing attachment, only variables that were significant were retained in the model, and the sex variable did not meet the threshold for inclusion in the model.

      The reviewer asks about the effect of adding social risk factor variables on the c-statistic. We provided that information in tables 7A and 7B on page 28 of the testing attachment, and we show the results for the dual eligible (DE) and low

AHRQ variables again, below, in Table 1, for your convenience. The c-statistic remains virtually unchanged after adding either of the two variables to the model, for both HOPDs and ASCs. We did not submit the results with the race variable (black), however the results are presented below in Table 1, along with the confidence intervals that were requested by another reviewer. Like the results with the dual eligible and low AHRQ variables, the c-statistic remains virtually unchanged after adding the race variable to the model. (Note that "black" is the only race variable available for our analyses in CMS claims data.) We also have the results for the race variable for the other social risk factor analyses we presented in the testing attachment, and have included this information along with the original results (for DE and AHRQ SES variables) in Attachment A. Overall, the results of these analyses for the race (black) variable are similar to the AHRQ SES results.

Table 1: C-statistics for model with and without social risk factors

Social risk factor	HOPDs	95% CI	ASCs	95% CI
None (reference)	0.684	0.682-0.687	0.653	0.650-0.656
DE	0.687	0.684-0.689	0.654	0.651-0.657
AHRQ SES	0.685	0.682-0.688	0.654	0.651-0.657
Race (black)	0.685	0.682-0.688	0.654	0.651-0.657

- **Issue 3: Social risk factor adjustment.** One methods panel reviewer stated that: *"The very high correlation between the scores with and without the factors can be used on either side of the decision to include– if it makes little difference, there is no harm in including the two factors, and there is probably not much harm created by excluding the two factors."*
    - **Developer Response 3:** We thank the reviewer for their comment. However, including the two factors creates the potential for harm since including social risk factors could remove incentives for providers to improve the quality of care for patients with social risk factors, and can send a signal that there are different standards of care for patients with and without social risk factors. Further, adjusting for social risk factors can mask any true differences between providers with respect to the quality of care that they deliver to vulnerable patients. On the other hand, risk adjustment for social risk factors may be appropriate when measures are sensitive to factors, providers have limited opportunity mitigate the additional risk, and not adjusting burdens providers caring for patients in underserved communities. While we have submitted other CMS measures with social risk factor adjustment to NQF, CMS has weighed the competing factors for this measure and decided the potential harm of adjusting the measures outweighs the benefit.
- **Issue 4: External empiric validity.** One Methods Panel reviewer stated: *"I think they should have attempted some analyses on the 'Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at ASCs (ASC General Surgery)' for facilities that have adequate volumes of target procedures."*
    - **Developer Response 4:** We thank the reviewer for this input. However, as stated in the testing form, this is not a viable approach. Many ASCs specialize in a single procedure (for example, in 2017, more than 60 percent of ASCs were single-specialty), and

gastroenterology is one of the most common single- specialty facility types.[o] Therefore, few ASCs performing colonoscopies are the same facilities that would be measured in the ASC General Surgery measure.

- **Issue 5: External empiric validity.** One Methods Panel reviewer stated that the developer "*could have done some validity testing on the outcome – what proportion of the numerator hospitalizations are related to the colonoscopy*? "
    - **Developer Response 5:** We've previously assessed and published this proportion. CORE examined the top reasons for return to the hospital, using an earlier version of the colonoscopy measure (the currently endorsed version), run on state HCUP data, and included these results in a manuscript we published on the measure.[p] The top reasons, shown below in Table 2 (excerpted from the manuscript), are either clearly related to the colonoscopy (such as laceration) or possibly related to having undergone a procedure requiring bowel preparation and anesthesia (such as atrial fibrillation).

Table 2. Top 10 Most Frequent Diagnoses Accompanying Unplanned Hospital Visits Within 7 Days of Outpatient Colonoscopy in the Study Cohort

ICD-9-CM code	Code description	% Of all unplanned hospital visits
998.11	Hemorrhage complicating a procedure	6.4
998.2	Accidental operative laceration	3.0
789	Abdominal pain unspecified site	3.0
578.9	Gastrointestinal hemorrhage, unspecified	2.7
786.5	Chest pain, unspecified	1.9
599	Urinary tract infection, unspecified	1.8
427.31	Atrial fibrillation	1.8
786.59	Other chest pain	1.7
486	Pneumonia, organism not otherwise specified	1.6
780.2	Syncope and collapse	1.5

[o] MedPAC's Report to Congress, Chapter 5, Ambulatory Surgical Center Services, March 2019. http://www.medpac.gov/docs/default-source/reports/mar19_medpac_ch5_sec.pdf?sfvrsn=0; Accessed December 9, 2019.

[p] Ranasinghe I, Parzynski CS, Searfoss R, Montague J, Lin Z, Allen J, Vender R, Bhat K, Ross JS, Bernheim S, Krumholz HM, Drye EE. Differences in Colonoscopy Quality Among Facilities: Development of a Post-Colonoscopy Risk-Standardized Rate of Unplanned Hospital Visits. Gastroenterology. Jan 2016;150(1):103-13.

- In addition, a 2018 single-center study examined the medical records (including medication information) of patients who experienced an emergency department (ED) visit within 7 days of an outpatient colonoscopy.[q] The study authors extracted patients' chief complaint from medical records, assigned the chief complaints as related or unrelated to the colonoscopy, and found that 68% of the reasons for the ED visit were due to the colonoscopy. The most common reasons for related ED visits were abdominal pain (38.2%), gastrointestinal bleeding (29.7%), cardiopulmonary disorders (12.7%), and nausea/vomiting (4.2%). The authors also identified a case of a vascular adverse event due to withholding anticoagulation for the procedure.

[q] Grossberg LB, Vodonos A, Papamichael K, Novack V, Sawhney M, Leffler DA. Predictors of post-colonoscopy emergency department use. Gastrointest Endosc. Feb 2018;87(2):517-525.