NATIONAL QUALITY FORUM
Driving measurable health improvements together

Scientific Methods Panel Discussion Guide

FALL 2020 EVALUATION CYCLE
October 28-29, 2020

This work is funded by the Centers for Medicare and Medicaid Services under contract HHSM-500-2017-00060I – 75FCMC19F0007.

Contents

Background

The Scientific Methods Panel provides NQF standing committees with evaluations of submitted complex measures' Scientific Acceptability (specifically, the "must-pass" subcriteria of reliability and validity), using NQF's standard measure evaluation criteria for new and maintenance measures.

This discussion guide contains details of the complex measures submitted for evaluation during the fall 2020 measure evaluation cycle. It also contains summaries of preliminary measure analyses and responses to these analyses composed by developers. The Scientific Methods Panel (SMP) utilizes this document during measure evaluation meetings to facilitate conversations between the Panel, measure developers, and NQF staff. This cycle, 25 complex measures were evaluated by the SMP. Four are up for discussion and revote. Five have been pulled by SMP members or NQF staff for further discussion, although they have passed NQF's Scientific Acceptability criterion.

After the SMP reviews measures, those that pass Scientific Acceptability move on to their respective standing committee for measure evaluation of the remaining NQF standard measure evaluation criteria (specifically, Importance to Measure and Report, Feasibility, Usability and Use, and requirements for Related and Competing Measures). Measures that do not pass the SMP can be pulled by a standing committee member for further discussion and re-vote if it is an eligible measure. Please refer to "Frequently Asked Questions" in NQF's standard measure evaluation criteria for details.

Measures for Discussion (Brief)

Subgroup 1

- 0505 Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Acute Myocardial Infarction (AMI) Hospitalization (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
  - **Reliability:** H-1; M-4; L-4; I-0 **Consensus Not Reached**
  - **Validity:** H-0; M-8; L-1; I-0 **Pass**
- 1891 Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Chronic Obstructive Pulmonary Disease (COPD) Hospitalization (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
  - **Reliability:** H-1; M-5; L-3; I-0 **Pass**
  - **Validity:** H-0; M-7; L-2; I-0 **Pass**
- 2515 Hospital 30-day, All-Cause, Unplanned, Risk-Standardized Readmission Rate (RSRR) Following Coronary Artery Bypass Graft (CABG) Surgery (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
  - **Reliability:** H-1; M-7; L-1; I-0 **Pass**
  - **Validity:** H-1; M-5; L-3; I-0 **Pass**

Subgroup 2

- 3599 Pediatric Asthma Emergency Department Use (University of California San Francisco)
  - **Reliability:** H-2; M-5; L-0; I-1 **Pass**
  - **Validity:** H-0; M-4; L-3; I-1 **Pass**
- 0230 Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
  - **Reliability:** H-0; M-5; L-3; I-0 **Pass**
  - **Validity:** H-0; M-6; L-1; I-1 **Pass**

Subgroup 3

- 3592 Global Malnutrition Composite Score (Avalere Health LLC)
  - **Reliability:** H-2; M-4; L-0; I-2 **Pass**
  - **Validity:** H-0; M-6; L-0; I-2 **Pass**
  - **Composite Construction:** H-2; M-3; L-2; I-1 **Pass**
- 0141 Patient Fall Rate (American Nurses Association)
  - **Reliability:** H-0; M-7; L-0; I-1 **Pass**
  - **Validity:** H-0; M-2; L-6; I-0 **Not Pass**
- 0202 Falls with Injury (American Nurses Association)
  - **Reliability:** H-0; M-7; L-0; I-1 **Pass**
  - **Validity:** H-0; M-1; L-5; I-2 **Not Pass**

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

Subgroup 1

- 0330 Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Heart Failure (HF) Hospitalization (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-0; M-7; L-1; I-0 **Pass**
    - **Validity:** H-2; M-5; L-1; I-0 **Pass**
- 0506 Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Pneumonia Hospitalization (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-1; M-7; L-1; I-0 **Pass**
    - **Validity:** H-0; M-8; L-1; I-0 **Pass**
- 2888 Risk-Standardized Acute Admission Rates for Patients with Multiple Chronic Conditions (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
    - **Reliability**: H-7; M-1; L-0; I-0 **Pass**
    - **Validity:** H-3; M-3; L-2; I-0 **Pass**
- 3597 Clinician-Group Risk-Standardized Acute Hospital Admission Rate for Patients with Multiple Chronic Conditions under the Merit-based Incentive Payment System (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-5; M-2; L-0; I-1 **Pass**
    - **Validity:** H-0; M-7; L-1; I-0 **Pass**
- 3596 Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Acute Ischemic Stroke Hospitalization with Claims-Based Risk Adjustment for Stroke Severity (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-3; M-5; L-0; I-0 **Pass**
    - **Validity:** H-1; M-5; L-2; I-0 **Pass**
- 2158 Medicare Spending Per Beneficiary (MSPB) - Hospital (Acumen, LLC / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-7; M-0; L-0; I-0 **Pass**
    - **Validity:** H-1; M-6; L-0; I-0 **Pass**

Subgroup 2

- 0229 Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following heart failure (HF) hospitalization (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-4; M-4; L-0; I-0 **Pass**
    - **Validity:** H-0; M-6; L-1; I-1 **Pass**
- 0531 Patient Safety and Adverse Events Composite (IMPAQ International / Centers for Medicare & Medicaid Services)
    - **Reliability:** -2; M-5; L-0; I-1 **Pass**
    - **Validity:** H-2; M-4; L-1; I-1 **Pass**

    - **Composite Construction:** H-2; M-4; L-1; I-1 **Pass**
- 0468 Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Pneumonia Hospitalization (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-4; M-4; L-0; I-0 **Pass**
    - **Validity:** H-1; M-5; L-1; I-1 **Pass**
- 1550 Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) (Centers for Medicare & Medicaid Services)
    - **Reliability:** H-2; M-6; L-0; I-0 **Pass**
    - **Validity:** H-0; M-6; L-1; I-1 **Pass**
- 1551 Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) (Centers for Medicare & Medicaid Services)
    - **Reliability:** H-2; M-5; L-1; I-0 **Pass**
    - **Validity:** H-0; M-7; L-0; I-1 **Pass**

Subgroup 3

- 3568 Person-Centered Primary Care Measure (Virginia Commonwealth University School of Medicine)
    - **Reliability**: H-2; M-3; L-1; I-2 **Pass**
    - **Validity:** H-0; M-6; L-0; I-2 **Pass**
- 3567 Hemodialysis Vascular Access: Practitioner Level Long-term Catheter Rate (University of Michigan Kidney Epidemiology and Cost Center)
    - **Reliability:** H-1; M-7; L-0; I-0 **Pass**
    - **Validity:** H-1; M-5; L-1; I-1 **Pass**
- 1623 Bereaved Family Survey (Department of Veterans Affairs / Hospice and Palliative Care)
    - **Reliability:** H-1; M-7; L-0; I-0 **Pass**
    - **Validity:** H-0; M-6; L-2; I-0 **Pass**
- 3235 Hospice and Palliative Care Composite Process Measure—Comprehensive Assessment at Admission (Abt Associates / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-5; M-3; L-0; I-0 **Pass**
    - **Validity:** H-2; M-5; L-1; I-0 **Pass**
    - **Composite Construction:** H-2; M-6; L-0; I-0 **Pass**
- 1893 Hospital 30-Day, all-cause, risk-standardized mortality rate (RSMR) following chronic obstructive pulmonary disease (COPD) hospitalization (Yale Center for Outcomes Research and Evaluation (CORE) / Centers for Medicare & Medicaid Services)
    - **Reliability:** H-0; M-6; L-1; I-0 **Pass**
    - **Validity:** H-2; M-5; L-0; I-0 **Pass**

Measures For Discussion (Detailed)

Subgroup 1

Measure #0505: Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Acute Myocardial Infarction (AMI) Hospitalization

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level 30-day, all-cause, RSRR for patients discharged from the hospital with a principal diagnosis of AMI. Readmission is defined as unplanned readmission for any cause within 30 days of the discharge date for the index admission. Readmissions are classified as planned and unplanned by applying the planned readmission algorithm. CMS annually reports the measure for patients who are 65 years or older and enrolled in fee-for-service (FFS) Medicare and hospitalized in non-federal hospitals or are patients hospitalized in Veterans Health Administration (VA) facilities.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk-adjustment:** Risk-adjusted for 31 risk factors; social risk factors (dual eligibility and ASPE SES index) were tested but not included in the final specification
- **Not based on a sample**
- **Ratings for reliability:** H-1; M-4; L-4 I-0 → Consensus not reached
    - Reliability testing was conducted at the measure score level:
        - The developer performed two types of reliability testing using Medicare Fee-For-Service Administrative Claims Data and VA Administrative data using a minimum case volume of 25
            - In the 65 years and older population, the developer estimated the overall measure score reliability by calculating the intraclass correlation coefficient (ICC) using a split sample (i.e. test-retest) method for a three-year period (2016-2019).
            - Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSRR for each hospital was 0.424
            - Estimated the facility-level reliability through signal-to-noise reliability using a minimum case volume of 25
            - The median reliability score was 0.51, ranging from 0.14 to 0.91. The 25th and 75th percentiles were 0.33 and 0.66, respectively
        - While results are within the lower range of moderate agreement based on the Landis and Koch standard, many SMP members felt the number was too low
    - **Ratings for validity:** H-0; M-8; L-1; I-0 → Measure passes with MODERATE rating
    - Validity testing conducted at the measure score level:
        - The developer examined the relationship of performance in the HF RSRRs with three external measures of hospital quality:
            - Hospital Star Rating readmission group score

        - The correlation between AMI RSRRs and Star-Rating readmissions score is -0.413, which suggests that hospitals with lower AMI RSRRs are more likely to have higher Star Rating Readmission scores
    - Overall Hospital Star Rating
        - The correlation between AMI RSRRs and Star Rating summary score is -0.266, which suggests that hospitals with lower AMI RSRRs are more likely to have higher Star Rating summary scores
    - AMI Excess Days in Acute Care (EDAC):
        - The correlation between AMI RSRRs and AMI EDAC scores is 0.425, which suggests that hospitals with lower AMI RSRRs are more likely to have lower AMI EDAC scores
- While a few members expressed concerns about the measure's ability to identify meaningful differences in performance and the risk adjustment model, overall, the SMP agreed that the tests showed moderate validity

ITEMS TO BE DISCUSSED

- **Action items:**
    - Discuss reliability testing/results and revote on reliability
    - Some SMP members expressed that the result of split-sample analysis and signal-to-noise analyses were low, barely crossing the lower limit of established moderate validity standards. Are the results sufficient enough to consider this measure moderately reliable?

Measure #1891: Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Chronic Obstructive Pulmonary Disease (COPD) Hospitalization Measure Title (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level 30-day, all-cause, RSRR for patients discharged from the hospital with either a principal discharge diagnosis of COPD or a principal discharge diagnosis of respiratory failure with a secondary diagnosis of acute exacerbation of COPD. The outcome (readmission) is defined as unplanned readmission for any cause within 30 days of the discharge date for the index admission (the admission included in the measure cohort). A specified set of planned readmissions do not count in the readmission outcome. CMS annually reports the measure for patients who are 65 years or older and are enrolled in fee-for-service (FFS) Medicare and hospitalized in non-federal hospitals or are patients hospitalized in Veterans Health Administration (VA) facilities.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk adjustment:** Risk-adjusted for 40 risk factors; social risk factors (dual eligibility and ASPE SES index) were tested but not included in the final specification
- **Ratings for reliability:** H-1; M-5; L-3; I-0 → Measure passes with MODERATE rating

- Reliability testing conducted at the measure score level:
    - The developer performed two types of reliability testing using Medicare Fee-For-Service Administrative Claims Data using a minimum case volume of 25
        - In the aged 65 years and older population, the developer estimated the overall measure score reliability by calculating the ICC using a split sample (i.e. test-retest) method for a three-year period (2016-2019)
            - Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSRR for each hospital was 0.406
        - Estimated the facility-level reliability through signal-to-noise reliability using a minimum case volume of 25
            - Under the results, the developer reported the median reliability score was 0.43, ranging from 0.11 to 0.90. The 25th and 75th percentiles were 0.25 and 0.60, respectively. The developer has confirmed the median score is 0.43
    - While the SMP agreed that the results from reliability tests showed moderate reliability, NQF staff are pulling this measure for discussion based on the review of other measures and the inconsistencies found in the submission form
- **Ratings for validity:** H-0; M-7; L-2; I-0 →Measure passes with MODERATE rating
- Validity testing conducted at the measure score level:
    - The developer examined the relationship of performance in the HF RSRRs with three external measures of hospital quality:
        - Hospital Star Rating readmission group score
            - The correlation between COPD RSRRs and Star Rating Readmissions score is -0.442, which suggests that hospitals with lower COPD RSRRs are more likely to have higher Star Rating Readmission scores
        - Overall Hospital Star Rating
            - The correlation between COPD RSRRs and Star Rating summary score is -0.286, which suggests that hospitals with lower COPD RSRRs are more likely to have higher Star Rating summary scores
- The developer also validated the measure through face validity using an 11-member Technical Expert Panel (TEP).
    - Of the 10 TEP members who responded to the developer's survey, 90% agreed (70% moderately or strongly agreed) that the measure will provide an accurate reflection of quality
- While a few members expressed concerns about the measure's ability to identify meaningful difference in performance and the risk adjustment model, overall, the SMP agreed that the tests showed moderate validity

ITEMS TO BE DISCUSSED

- Clarification from the developer needed on the results of the signal-to-noise test.

- **Action items:**

Some SMP members expressed that the result of split-sample analysis and signal-to-noise analyses were low, barely crossing the lower limit of established moderate validity standards. Are the results sufficient enough to consider this measure moderately reliable?

Measure #2515: Hospital 30-day, All-Cause, Unplanned, Risk-Standardized Readmission Rate (RSRR) Following Coronary Artery Bypass Graft (CABG) Surgery Measure Title (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level RSRR, defined as unplanned readmission for any cause within 30-days from the date of discharge for a qualifying index CABG procedure. The measure was developed using Medicare FFS patients 65 years and older and was tested in all-payer patients 18 years and older. An index admission is the hospitalization for a qualifying isolated CABG procedure considered for the readmission outcome.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data
- **Level of analysis:** Facility
- **Risk adjustment:** Risk-adjusted for 26 risk factors; social risk factors (dual eligibility and ASPE SES index) were tested but not included in the final specification
- **Not based on a sample**
    - **Ratings for reliability:** H-1; M-7; L-1; I-0 → Measure passes with MODERATE rating
    - Reliability testing conducted at the measure score level:
        - The developer performed two types of reliability testing using Medicare Fee-For-Service Administrative Claims Data using a minimum case volume of 25
            - In the aged 65 years and older population, the developer estimated the overall measure score reliability by calculating the ICC using a split sample (i.e. test-retest) method for a three-year period (2016-2019)
                - Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSRR for each hospital was 0.436
            - Estimated the facility-level reliability through signal-to-noise reliability using a minimum case volume of 25
                - The median reliability score was 0.60, ranging from 0.27 to 0.92. The 25th and 75th percentiles were 0.45 and 0.71, respectively
            - The SMP agreed that the results from reliability tests showed moderate reliability but experienced concerns about reliability in lower volume facilities
    - **Ratings for validity:** H-1; M-5; L-3; I-0 → Measure passes with MODERATE
    - Validity testing conducted at the measure score level:
        - The developer examined the relationship of performance in the HF RSRRs with three external measures of hospital quality:
            - Hospital Star Rating readmission group score

        - The correlation between COPD RSRRs and Star Rating Readmissions score is -0.442, which suggests that hospitals with lower COPD RSRRs are more likely to have higher Star Rating Readmission scores
    - Overall Hospital Star Rating
        - The correlation between COPD RSRRs and Star Rating summary score is -0.286, which suggests that hospitals with lower COPD RSRRs are more likely to have higher Star Rating summary scores
- The developer also validated the measure through face validity using an 11-member TEP.
    - Of the 10 TEP members who responded to the developer's survey, 90% agreed (70% moderately or strongly agreed) that the measure will provide an accurate reflection of quality
- While a few members expressed concerns about the measure's ability to identify meaningful difference in performance and the risk adjustment model, overall, the SMP agreed that the tests showed moderate validity

ITEMS TO BE DISCUSSED

- **Action items:**

SMP member asked to pull this measure to discuss How do panel members view their rational behind a high/moderate vote or a low/insufficient vote for similar reliability statistics?

Subgroup 2

Measure #3599: Pediatric Asthma Emergency Department Use

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This measure estimates the rate of emergency department visits for children ages three to 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 one, 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:** Six risk factors included with three social risk factors – Age, Gender, Chronic condition indicator, % households below the poverty level, % population with less than high school education, % male unemployment for 25- to 60-year olds
- **Not based on a sample**
    - **Ratings for reliability:** H-2; M-5; L-0; I-1 → Measure passes with MODERATE rating
    - Reliability testing conducted at the score level:
        - Developer tested the measure using split-sample analysis and ICC calculations for score level reliability testing in 26 health plans in Massachusetts and 101 health plans in California
            - MA health plans ICC: 0.72
            - CA health plans ICC: 0.86
        - Testing was conducted using a risk-adjusted approach in a mixed effect model
        - While it was not clear which ICC was used, SMP members agreed with this approach and considered the results to be moderate to strong.
        - One SMP member called for greater clarity regarding the definition of the denominator
    - **Ratings for validity:** H-0; M-4; L-3; I-1 → Consensus not reached
    - Validity testing conducted at the score level:
        - Developer tested the measure for construct validity by using predicted performance for the plan-level random effect in the risk adjustment models and then transformed that into a Z-score. Pairwise correlations were made to select HEDIS measures. Correlation results:
            - Medication Management in Asthma Compliance 50% CC: 0.12
            - Medication Management in Asthma Compliance 75% CC: 0.13
            - Child Vaccines: 0.33
            - ACE Monitoring: 0.05
            - Cervical Cancer Screening: 0.04
            - Low Back Pain Imaging: 0.05
        - Predictive validity was conducted as a secondary analysis at the clinic level in Vermont, assessing a quality innovation (QI) learning collaborative reduction in emergency department (ED) utilization through a difference in difference analysis
            - Adjusted marginal ED visit rates were superior in QI participants
                - Non participants change over time was 1.58 visits per 100 person-years
                - Participants change over time was -6.28 visits per 100 person-years
                - Difference in differences was -7.28
            - This suggests that the measure is responsive to QI initiatives related to ED utilization reduction for asthma
        - SMP identified threats to validity:
            - The exclusion of testing only in data from MA.
            - Missing data: Certain elements have high level of missingness, which may partially account for the difference in model performance between Medicaid and all payer data

- Risk adjustment: R2 highly variable between development and validation sets. Confidence intervals and p values not given. Very few variables (6) included in the model. Method of variable selection was a priori and not well explained
- "Given that secondary diagnosis for asthma may be unrelated to the reason for ER visit/admission, the validity of the measure as constituted has not been established."
- "Low construct validity values."
- "Concerns remain with the number of plans that have a relative high percentage of members who are missing social risk factor data. The social risk factors are key adjustment variables in the risk adjustment models. 20,000 of 85,000 members are in plans for which 10% of more of members are missing social risk data. And concerns remain with the risk-adjustment model. Given the possible differences in SES factors for APCD and Medicaid populations, applying a singular risk model to all populations may prove challenging/problematic."

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- **Action items:**
  - Resolve consensus not reached vote on validity through discussion and revote.

Measure #0230: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level 30-day RSMR for patients discharged from the hospital with a principal diagnosis of AMI. Mortality is defined as death for any cause within 30 days after the date of admission for the index admission. CMS annually reports the measure for patients who are 65 years or older and are either Medicare FFS beneficiaries and hospitalized in non-federal hospitals or are hospitalized in VA facilities.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk-adjusted:** 27 risk factors; social risk factors were tested but not included
- **Not based on a sample**
- **Ratings for reliability:** H-0; M-5; L-3; I-0 → Measure passes with MODERATE rating
  - Reliability testing conducted at the measure score level by calculating the ICC using a split sample (i.e. test-retest) method and estimating the facility-level reliability (signal-to-noise reliability)
  - The ICC for hospitals with 25 admissions or more was 0.428
  - The median reliability score was 0.59 (ranging from 0.20 to 0.93) for the signal-to-noise testing for each hospital with at least 25 admissions
  - Most SMP members agree the reliability tests are appropriate but the results show moderate reliability

  - A couple of members voiced concerns that signal-to-noise Ratio (SNR) results suggest below acceptable level of reliability at the score level (<0.7), thus there is low certainty that the performance measure scores are reliable. The 75th percentile of SNR reliability estimate was 0.72, suggesting that about 70-75% of providers assessed did not meet the 0.7 criteria
  - One asked for clarification of how the 25-case threshold was established
  - One expressed disagreement of using the Landis modifiers in the split sample testing
- **Ratings for validity:** H-0; M-6; L-1; I-1 → Measure passes with MODERATE rating
  - Validity testing conducted at the performance measure score level, including both empirical validity testing (by comparing to CMS star ratings mortality scores and star rating summary scores), and systematic assessment of face validity
  - The correlation between AMI RSMRs and Star-Rating mortality score is -0.409, which suggests that hospitals with lower AMI RSMRs are more likely to have higher Star-Rating mortality scores
  - The correlation between AMI RSMRs and Star-Rating summary score is -0.204, which suggests that hospitals with lower AMI RSMRs are more likely to have higher Star-Rating summary scores
  - The median absolute change in hospitals' RSMRs when adding a dual eligibility indicator is 0.07% (interquartile range [IQR] -0.005% – 0.009%) with a correlation coefficient between RSMRs for each hospital with and without dual eligibility added of 0.999. The median absolute change in hospitals' RSMRs when adding a low Agency for Healthcare Research and Quality (AHRQ) SES Index score indicator to the model is 0.049% (IQR -0.021% – 0.068%) with a correlation coefficient between RSMRs for each hospital with and without an indicator for a low AHRQ SES Index score adjusted for cost of living at the census block group level is 0.978.
  - The addition of any of these variables into the hierarchical model has little to no effect on hospital performance (c-statistic remains 0.73)
  - Most member think the validity tests are appropriate, the results are moderate, and the exclusions are appropriate
  - One member asked for clarification on whether patients with MI died in the hospital are included in the numerator and whether this measure has been compared to in-hospital mortality due to MI
  - One member voiced concern about this measure performance being topped out, based on the fact that 98% [2284/(2284+28+16)] of hospitals assessed performed no different from the U.S. national rate
  - A couple of members asked for clarification on how model calibration is kept updated using newer datasets
  - Most members thought the risk adjustment model is solid, but a few members had questions about not including risk factors in the model because of no added predictive power and no change in hospital performance rankings. It would be useful to know the rate of hospitals that would have change rank if social-risk factors would have been included, and the rationale explaining why including other risk factors with nonsignificant coefficients did not apply to social risk factors
  - One member noted that "In 2020, dual eligibility and the AHRQ SES index have effect sizes (odds ratios) of 1.08 and 1.07 when added independently to the model" is contradictory to the statement of "the relationship between dual-eligible status and AHRQ low SES is in the opposite direction."

  - One asked for clarification on whether clinical plausibility of risk factors included in the model were assessed

- *ITEMS TO BE DISCUSSED*
- Additional clarifying information from the developer
- **Action items:**
  - Regarding reliability being low for half of the entities, can the developers comment of the impact of the observed reliability on misclassification or other consequences?
  - Clarification on how model calibration is kept updated using newer datasets
  - When coefficients are nonsignificant, should the same inclusion/exclusion rationale apply to both social risk factors and clinical risk factors?

Subgroup 3

Measure #3592: Global Malnutrition Composite Score (Composite) (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This composite measure of optimal malnutrition care focuses on adults 65 years and older admitted to inpatient service who received care appropriate to their level of malnutrition risk and/or malnutrition diagnosis if properly identified. Best practices for malnutrition care recommend adult inpatients to be screened for malnutrition risk, assessed to confirm findings of malnutrition if found at risk and have the proper severity of malnutrition indicated along with a corresponding nutrition care plan that addresses the respective severity of malnutrition.
  The malnutrition composite measure includes four component measures which are first scored separately. The overall composite score is derived from averaging the individual performance scores.
  1. Screening for malnutrition risk at admission.
  2. Completing a nutrition assessment for patients who screened for risk of malnutrition.
  3. Appropriate documentation of malnutrition diagnosis in the patient's medical record if indicated by the assessment findings.
  4. Development of a nutrition care plan for malnourished patients including the recommended treatment plan.
  These four measures represent the key processes of care and generated markers of malnutrition associated with the risk identification, diagnosis, and treatment of malnutrition in older hospitalized adults as supported by clinical guidelines.
- **Type of measure:** Composite
- **Data source:** Electronic Health Records
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: To meet minimum requirements for measure implementation in quality reporting programs, there must be a minimum of 20 eligible encounters per reporting entity for valid performance measure scoring.
  - **Ratings for reliability:** H-2; M-4; L-0; I-2 → Measure passes with MODERATE rating

    - Reliability testing conducted at the measure score level to calculate the ICC:
    - With case minimums, the ICC calculated was 0.839, and without case minimums, it resulted in an ICC of 0.647
    - One reviewer noted that it is unclear if the measure is limited to hospitals >= 50 infants. If not, then the reliability testing does not match the specifications
    - Reviewers generally agreed with the analysis, but two questioned whether the increased reliability since last submission was due to change in sample size, which is not part of the specifications
    - One reviewer suggested an alternate calculation for ICC at the health system level to yield more accurate result for ICC would be B/(B+W/n) where n is the number of sites per system and W/n is the variance of the average of scores across the n sites
- **Ratings for validity:** H-0; M-6; L-0; I-2 → Measure passes with MODERATE rating
    - Empirical validity testing conducted at both the measure score and data element level:
    - Reviewers generally agreed with the analysis
    - Empirical testing of the construct validity of the overall composite measure at the score level was conducted. A hierarchical linear regression model was used to demonstrate that the predictability of the model significantly improved when the components in aggregate were included into the model over standard predictors of these outcomes such as patient characteristics, primary diagnoses, and comorbidities
    - Construct validity of the critical data elements for the individual measure components was tested by developing a generalized linear (logistic) regression model
    - One reviewer specifically mentioned that the results presented contain an error (the text and chart about readmission contradict each other). The one finding about better measure performance being associated with lower readmission rates would seem to be evidence for validity, but the text says one thing and the graphic says the opposite.
- **Ratings for Composite:** H-2; M-3; L-2; I-1 → Measure passes with MODERATE rating

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- **Action items:**
    - Two SMP members voted insufficient on validity and asked to discuss the validity criterion.

Measure #0141: Patient Fall Rate

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** All documented falls, with or without injury, experienced by patients on eligible unit types in a calendar quarter. Reported as Total Falls per 1,000 Patient Days.
  (Total number of falls / Patient days) X 1000

Measure focus is safety.
Target population is adult acute care inpatient and adult rehabilitation patients.

- **Type of measure:** Outcome
- **Data source:** Electronic Health Records, Other, Paper Medical Records
- **Level of analysis:** Facility, Other
- **Not Risk-adjusted but stratification is used**
- **Not based on a sample**
    - **Ratings for reliability:** H-0; M-7; L-0; I-1 → Measure passes with MODERATE rating
    - Reliability testing conducted at the measure score level:
        - Reliability testing was done at the nursing care unit level and included: 1) signal-to-noise analysis; and 2) interclass correlation
            - Signal-to-noise analysis ranged from 0.66 (critical care units) to 0.83 (rehabilitation units)
            - Intraclass correlations ranged from 0.58 (critical care units) to 0.80
        - Reliability testing was also done at the hospital level and included: 1) squared correlation between the hospital score and the estimates true fall rate; 2) signal-to-noise; and 3) the association between the bootstrap hospital scores and true fall rate with Spearman's correlations
            - The average squared correlation between the hospital score and hospital fall rate was 0.69 ± 0.20 and ranged from 0.07-0.96 across hospitals
            - The average reliability score on the second signal-to-noise measure had mean 0.73 ± 0.18 and ranged from 0.06-0.99
            - The average hospital score rank was strongly associated with the rank of the true hospital fall rate (r = 0.84)

    - **Ratings for validity:** H-0; M-2; L-6; I-0 → Measure does not pass with LOW rating
    - Validity testing conducted at the measure score level:
        - Validity testing included: 1) convergent validity by exploring the relationship between the falls measure and other measures, such as nurse staffing, registered nurse (RN)-rated quality of care, and missed care, and 2) prior validity studies using vignettes to determine sensitivity and specificity of the measure
        - Developers reported "small to moderate significant ($p < 0.01$) negative correlations (ranging from -0.11 to -0.21) were found between RN reports of appropriate staffing and total fall rate for the hospital measure, and for all unit types except critical care. Moderate significant (ranging from -0.17 to -0.21) negative relationships between RN rated quality of care and total fall rates were found between RN reported quality of care and fall rates at the hospital level as well as for step-down, medical, and medical-surgical combined units. Ratings of missed care activities do not have any substantial relationships except some small correlations on step-down and medical units and at the hospital level."

        - SMP members raised concern that the correlations seem low and may not be the correct measures. There were also concerns with respect to the negative associations found for some of the measures
    - The developer also reported that sensitivity and specificity for identifying falls based on clinical vignettes did not vary substantially across providers
    - Threats to validity:
        - Exclusions: Some SMP members raised concerns regarding the lack of exclusions, stating that "the developer stated no exclusions, but certain nursing units are not included in this measure, and that "the MIF notes exclusions. However, the testing form notes 'no exclusions.' Thus, there's no rationale as to the exclusions & no analysis of them."
        - Meaningful differences: The developer states, "Due to within-hospital variability (resulting from true differences among unit fall rates, as well as randomness inherent in fall counts), we would expect a 'medium' difference in true hospital fall rates in a given year to be detectable about half the time."
            - Some SMP members found this to be concerning and added that the developers "do not specify what a 'medium' difference is and how that relates to the concept of 'meaningful."
        - The developer did not risk adjust this measure but stratified by six risk categories according to unit types
        - Some SMP members raised significant concern regarding the lack of risk adjustment. There was also concern regarding the approach to social risk factor (SRF) assessment
            - "…[On the surface the lack of risk adjustment seems problematic" and question the developer's rationale
            - "Only did a literature review to evaluate whether there should be SRF, which are not currently in the literature. Suggest doing an evaluation of National Database of Nursing Quality Indicators (NDNQI) data if possible to determine if SRF can be correlated to hospital falls."

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- **Action items:**
    - *Reliability – The SMP should discuss the following:*
        - One SMP member noted that while, overall, the reliability results look acceptable, the lower ranges the squared correlations and signal-'to-noise results "seem concerning as these numbers are exceptionally low
    - *Validity - The SMP should discuss the following:*
        - Concerns with the correlations being low with some being negative
        - Concerns related to the lack of risk adjustment and stratification only

- Concerns regarding the lack of exclusions
- Issues regarding the ability to identify meaningful differences in performance, as there were concerns raised with respect to the "medium" differences reported by the developer

Measure #0202: Falls with Injury

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** All documented patient falls with an injury level of minor or greater on eligible unit types in a calendar quarter. Reported as Injury falls per 1000 Patient Days.
  (Total number of injury falls / Patient days) X 1000
  Measure focus is safety.
  Target population is adult acute-care inpatient and adult-rehabilitation patients.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Records, Other, Paper Medical Records
- **Level of analysis:** Facility, Other
- **Not Risk-adjusted, but Stratified**
- **Not based on a sample**
  - **Ratings for reliability:** H-0; M-7; L-0; I-1 → Measure passes with MODERATE rating
  - Reliability testing conducted at the measure score level:
    - Reliability testing was done at the nursing care unit level and included signal-to-noise analysis and interclass correlation
      - Signal-to-noise analysis ranged from 0.38 (surgical units) to 0.70 (rehabilitation units).
      - Intraclass correlations ranged from 0.25 (critical care units) to 0.63 (rehabilitation units)
    - Reliability testing was also done at the hospital level and included the following: 1) squared correlation between the hospital score and the estimates true fall rate; 2) signal-to-noise; and 3) the association between the bootstrap hospital scores and true fall rate with Spearman's correlations
      - The average squared correlation between the hospital score and hospital fall rate was 0.66 ± 0.21 and ranged from 0.05-0.96 across hospitals
      - The average reliability score on the second signal-to-noise measure had mean 0.74 ± 0.21 and ranged from 0.04-0.99
    - The developer stated that the average "hospital score percentile rank was strongly associated with the rank of the true hospital injury fall rate (r = 0.96)"
  - **Ratings for validity:** H-0; M-1; L-5; I-2 → Measure does not pass with LOW rating
  - Validity testing conducted at the measure score level:
    - Validity testing included: 1) convergent validity by exploring the relationship between the falls measure and other measures, such as - percent of falls that were unassisted, nurse staffing, RN-rated quality of care, and missed care, 2) prior validity studies using vignettes to

determine sensitivity and specificity of the measure, and 3) Confirmatory Factor Analysis (CFA)

- Developers reported, "for the convergent validity tests, the link between unassisted fall rate and falls with injury was strong (0.48-0.58) for all unit types and the overall hospital measure. The correlations were small to moderate, for appropriate staffing and quality of care rating across most unit types."
    - SMP members raised concern that the correlations seem low.
- The developer also reported "the findings from the vignette study confirm that the 10 vignettes from the injury falls survey has resulted in latent structures that are appropriate for predicting the severity of the injury falls and provide additional evidence of both construct and convergent validity"
- For the confirmatory analysis, the developer stated, "The results from the initial CFA procedure did not indicate a good model fit (CFI = 0.883, TLI = 0.878, RMSEA = 0.06). In order to improve the fit of the measurement model, we repeated the CFA procedure by removing one item from the model. The final CFA assessment confirmed a good model fit and our hypothesis that a relationship between the 11 observed items and the two underlying latent factors exist (CFI = 0.95, TLI = 0.945, RMSEA = 0.042)."
- Threats to validity:
    - Exclusions: Some SMP members raised concerns regarding the lack of exclusions, stating that "the developer stated no exclusions, but certain nursing units are not included in this measure, and that "the MIF notes exclusions. However, the testing form notes 'no exclusions.' Thus, there's no rationale as to the exclusions & no analysis of them."
    - Meaningful differences: The developer states "Due to within-hospital variability (resulting from true differences among unit fall rates, as well as randomness inherent in fall counts), we would expect a 'medium' difference in true hospital fall rates in a given year to be detectable about half the time."
        - Some SMP members found this to be concerning and added that the developers "do not specify what a 'medium' difference is and how that relates to the concept of 'meaningful.'
    - The developer did not risk adjust this measure, but they stratified it by six risk categories according to unit types.
        - Some SMP members raised significant concern regarding the lack of risk adjustment. There was also concern regarding the approach to social risk factor (SRF) assessment. ("…on the surface the lack of risk adjustment seems problematic" and question the developer's rationale)
        - "Only did a literature review to evaluate whether there should be SRF, which are not currently in the literature.

Suggest doing an evaluation of NDNQI data if possible, to determine if SRF can be correlated to hospital falls."

ITEMS TO BE DISCUSSED

- The developer checked the box of conducting face validity, but no methods or results are reported
- **Action items:** Similar comments to #0141
    - *Reliability – The SMP should discuss the following:*
        - Some concern related to low reliability results for surgical units and the low range of reliability results for hospitals
    - *Validity - The SMP should discuss the following:*
        - Concerns with the correlations being low
        - Concerns related to the lack of risk adjustment and stratification only
        - Concerns regarding the lack of exclusions
- Issues regarding the ability to identify meaningful differences in performance, as there were concerns raised with respect to the "medium" differences reported by the developer

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

Subgroup 1 – Measure Details

Measure #0330: Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Heart Failure (HF) Hospitalization

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level RSRR for patients discharged from the hospital with a principal diagnosis of HF. Readmission is defined as unplanned readmission for any cause within 30 days of the discharge date for the index admission. Readmissions are classified as planned and unplanned by applying the planned readmission algorithm. CMS annually reports the measure for patients who are 65 years or older and are enrolled in FFS Medicare and hospitalized in non-federal hospitals or are patients hospitalized in VA facilities.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk-adjustment:** Risk-adjusted for 37 risk factors; social risk factors (dual eligibility and ASPE SES index) were tested but not included in the final specification
- **Not based on a sample**
    - **Ratings for reliability:** H-0; M-7; L-1; I-0 → Measure passes with MODERATE rating
        - Reliability testing conducted at the measure score level:
            - The developer performed two types of reliability testing using Medicare Fee-For-Service Administrative Claims Data and VA Administrative data using a minimum case volume of 25.
            - In the aged 65 years and older population, the developer estimated the overall measure score reliability by calculating the ICC using a split sample (i.e. test-retest) method for a three-year period (2016-2019).
            - Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSRR for each hospital was 0.587
            - Estimated the facility-level reliability through signal-to-noise reliability using a minimum case volume of 25
            - The median reliability score was 0.57, ranging from 0.14 to 0.96. The 25th and 75th percentiles were 0.31 and 0.75, respectively.
        - The SMP agreed that the results from reliability tests showed moderate reliability but experienced concerns about reliability in lower volume facilities.
    - **Ratings for validity:** H-2; M-5; L-1; I-0 → Measure passes with MODERATE rating
        - Validity testing conducted at the measure score level:
            - The developer examined the relationship of performance in the HF RSRRs with three external measures of hospital quality:

- Hospital Star Rating readmission group score: The correlation between HF RSRRs and Star-Rating readmissions score is -0.585, which suggests that hospitals with lower HF RSRRs are more likely to have higher Star-Rating readmission scores
- Overall Hospital Star Rating: The correlation between HF RSRRs and Star-Rating summary score is -0.378, which suggests that hospitals with lower HF RSRRs are more likely to have higher Star-Rating summary scores
- HF Excess Days in Acute Care (EDAC): The correlation between HF RSRRs and HF EDAC scores is 0.574, which suggests that hospitals with lower HF RSRRs are more likely to have lower HF EDAC scores

- The developer also validated the HF readmission administrative model against a medical record model with the same cohort of patients for which hospital-level HF readmission medical record data are available using a sample of 64,329 patients
  - The areas under the receiver operating characteristic (ROC) curve were 0.61 and 0.58, respectively, for the two models
  - The correlation coefficient of the standardized rates from the administrative and medical record models was 0.97
- While a few members expressed concerns about the measure's ability to identify meaningful difference in performance and the risk adjustment model, overall, the SMP agreed that the tests showed moderate validity

Measure #0506 Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Pneumonia Hospitalization

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level 30-day, all-cause, RSRR for patients discharged from the hospital with either a principal discharge diagnosis of pneumonia (including aspiration pneumonia) or a principal discharge diagnosis of sepsis (not severe sepsis) with a secondary diagnosis of pneumonia (including aspiration pneumonia) coded as present on admission (POA). Readmission is defined as an unplanned readmission for any cause within 30 days of the discharge date for the index admission. Readmissions are classified as planned and unplanned by applying the planned readmission algorithm. CMS annually reports the measure for patients who are 65 years or older and enrolled in FFS Medicare and hospitalized in non-federal hospitals or are patients hospitalized in VA facilities.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk adjustment:** Risk adjusted for 41 risk factors; social risk factors (dual eligibility and ASPE SES index) were tested but not included in the final specification
- **Not based on a sample**
- **Ratings for reliability:** H-1; M-7; L-1; I-0 → Measure passes with MODERATE rating

- Reliability testing conducted at the measure score level:
    - The developer performed two types of reliability testing using Medicare Fee-For-Service Administrative Claims Data and VA Administrative data using a minimum case volume of 25.
        - In the aged 65 years and older population, the developer estimated the overall measure score reliability by calculating the ICC using a split sample (i.e. test-retest) method for a three-year period (2016-2019)
            - Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSRR for each hospital was 0.544
        - Estimated the facility-level reliability through signal-to-noise reliability using a minimum case volume of 25
            - The median reliability score was 0.56, ranging from 0.13 to 0.96. The 25th and 75th percentiles were 0.34 and 0.73, respectively
    - The SMP agreed that the results from reliability tests showed moderate reliability but experienced concerns about reliability in lower volume facilities
- **Ratings for validity:** H-0; M-8; L-1; I-0 → Measure passes with MODERATE rating Validity testing conducted at the measure score level:
    - The developer examined the relationship of performance in the HF RSRRs with three external measures of hospital quality:
        - Hospital Star Rating readmission group score
            - The correlation between Pneumonia RSRRs and Star-Rating readmissions score is -0.564, which suggests that hospitals with lower Pneumonia RSRRs are more likely to have higher Star-Rating readmission scores
        - Overall Hospital Star Rating
            - The correlation between Pneumonia RSRRs and Star-Rating summary score is -0.371, which suggests that hospitals with lower Pneumonia RSRRs are more likely to have higher Star-Rating summary scores
        - Pneumonia Excess Days in Acute Care (EDAC):
            - The correlation between Pneumonia RSRRs and Pneumonia EDAC scores is 0.625, which suggests that hospitals with lower Pneumonia RSRRs are more likely to have lower Pneumonia EDAC scores
    - While a few members expressed concerns about the measure's ability to identify meaningful difference in performance and the risk adjustment model, overall, the SMP agreed that the tests showed moderate validity

Measure #2888: Risk-Standardized Acute Admission Rates for Patients with Multiple Chronic Conditions

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** Rate of risk-standardized acute, unplanned hospital admissions among Medicare FFS beneficiaries 65 years and older with multiple chronic conditions (MCCs) who are assigned to an Accountable Care Organization (ACO).
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Integrated Delivery System
- **Risk-adjusted:** Yes, risk adjusted and SES included
- **Ratings for reliability:** H-7; M-1; L-0; I-0 → Measure passes with HIGH rating
    - Reliability testing conducted at the measure score level:
        - The median signal-to-noise reliability was 0.96 for all ACO's with at least one attributed MCC patient (N=559) with an interquartile range of 0.94 to 0.98, calculated using one year of data. A split-half analysis is not provided. It should be noted that although the range was 0.12 to 0.99, the mean was 0.95 with a standard deviation of 0.05
        - The SMP did not raise any major concerns with reliability testing.
    - **Ratings for validity:** H-3; M-3; L-2; I-0 → Consensus not reached
    - Validity testing conducted at the measure score level:
        - The developer conducted face validity and empirical validity testing
        - For face validity, a TEP was convened a provided expert panel input as to the conditions, groupings, and modeling. Public commenting was also requested. Quantitative analyses for face validity was not conducted.
        - Empirical validity testing evaluated whether performance on the ACO measure was correlated with performance on five other ACO measures that assessed the same domains of quality (i.e., care coordination and management of chronic conditions): ACO1 - CAHPS *Getting Timely Care, Appointments, and Information*; ACO4 - CAHPS *Access to Specialists*; ACO8 - *Risk Standardized, All Condition Readmission; ACO27 - Diabetes: Hemoglobin A1c (HbA1c) Poor Control (>9%*); ACO28 - *Controlling High Blood Pressure*
            - There was little to no correlation between the ACO measure and the CAHPS measures
            - There was moderate correlation between the readmissions measure (spearman correlation 0.42, p<.001) as expected
            - There was weak correlation between the diabetes poor control measure (0.18, p<.001) and slightly negative but insignificant correlation with the control of high blood pressure measure (-0.07, p=0.673)
            - Some SMP members raised concern with the results of the validity testing: noting that "4 of the 5 comparator measures hypothesized a weak or poor relationship with the measure" and "two measures can be uncorrelated because they have no conceivable relationship to each other

- Risk adjustment: The risk adjustment model utilized 49 variables, demographic (age), 46 clinical (diagnosis groupers and functional status), and two social risk variables (SES index and specialist density)
    - Some SMP members raised concern regarding the model fit, stating that, "the model was evaluated using deviance R-squared, which was 0.111 indicating the model explains 11.1% of variation in admission rates. This is NOT [an] indication of a very strong model"

Measure #3597: Clinician-Group Risk-Standardized Acute Hospital Admission Rate for Patients with Multiple Chronic Conditions under the Merit-based Incentive Payment System

MEASURE HIGHLIGHTS

- New Measure
- **Description:** Risk-Standardized rate of acute, unplanned hospital admissions among Medicare FFS patients aged 65 years and older with MCCs.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Yes, 49 risk factors adjusted and includes SES factors (CMS has decided to adjust for the AHRQ SES Index and physician-specialist density)
- **Not based on a sample**
    - **Ratings for reliability:** H-5; M-2; L-0; I-1 → Measure passes with HIGH rating
    - Reliability testing conducted at the measure score level:
        - Signal-to-noise and intraclass correlation coefficient: Developers examined the distribution of mean and median reliabilities across patient volume for these clinician groups and identified a case minimum of 18 MCC patients per clinician group as providing adequate reliability. Using this case minimum and the group size threshold of >15 clinicians per group, mean and median reliability for 4,044 TINs was 0.809 and 0.873, respectively (range 0.413-0.999, IQR 0.683-0.961)
        - The SMP did not raised any major concerns with reliability testing
    - **Ratings for validity:** H-0; M-7; L-1; I-0 → Measure passes with MODERATE rating
    - Validity testing conducted at the measure score level:
        - New measure → only face validity was conducted
            - The developer convened a TEP to provide input as to the conditions, groupings, and modeling. Public commenting was also requested. A survey of the TEP showed 83% of respondents agreed that the MIPS MCC admission measure can be used to distinguish good from poor quality of care
            - Of 11 member assessing ability to distinguish good from poor, five of 11 somewhat agreed, five moderately agreed, and one strongly disagreed.

- One SMP member stated that, "Face validity testing should be both transparent and systematic. I do not think the process used meets those criteria."
- However, other SMP members did not raise any major concern

- Threats to validity: The final patient-level risk-adjustment model included 49 variables (47 demographic and clinical variables and two social risk factors). They used a negative binomial regression model with linear variance (NB-1) to risk adjust the measure. The model built off work done for the ACO MCC admission measure. Social risk factors included low AHRQ SES index and low physician-specialist density

Measure #3596: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Acute Ischemic Stroke Hospitalization with Claims-Based Risk Adjustment for Stroke Severity

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The measure estimates the hospital-level, RSMR for patients discharged from the hospital with a principal discharge diagnosis of acute ischemic stroke. The outcome is all-cause, 30-day mortality, defined as death from any cause within 30 days of the index admission date, including in-hospital death, for stroke patients. This is a re-specified measure with a cohort and outcome that is harmonized with CMS's current publicly reported claims-based stroke mortality measure and includes the National Institutes of Health (NIH) Stroke Scale as an assessment of stroke severity upon admission in the risk-adjustment model. This measure uses Medicare FFS administrative claims for the cohort derivation, outcome, and risk adjustment.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data
- **Level of analysis:** Facility
- **Risk-adjusted:** 20 risk factors; social risk factors (dual eligibility and ASPE SES index) were tested but not included in the final specification
- **Not based on a sample**
- **Ratings for reliability:** H-3; M-5; L-0; I-0 →Measure passes with MODERATE rating
  - Reliability testing conducted at the performance measure score level
  - The median signal-to-noise reliability score for all hospitals in the testing sample (N=329) and hospitals with at least 25 cases (N=292) was 0.72, ranging from 0.01 to 0.95.
  - The SMP reviewers all agree the reliability test was appropriate and the result showed high reliability and clinically meaningful difference with some concern about reliability in low-volume hospitals
- **Ratings for validity:** H-1; M-5; L-2; I-0 →Measure passes with MODERATE rating
  - Validity testing was conducted at both the data element level and performance measure score level by empirical validity testing and systematic assessment of face validity
  - Data Element Validity using GWTG-Stroke Registry: When comparing the NIH Stoke Scale scores within the GWTG-Stroke Registry and administrative claims data, 93% of the scores from the two data sources are within five points of each other and 84% are within two points. The distributions of NIH Stroke Scale scores from the administrative

and GWTG-Stroke Registry data are similar with a Pearson Correlation Coefficients: 0.993 and a weighted kappa of 0.842
- Measure score level testing: the correlation between stroke RSMRs and the Overall Star Ratings Mortality scores is 0.422
- The model was evaluated using the c-statistic, which was 0.86 indicating strong predictive ability
- The face validity testing results demonstrated working group agreement with overall face validity of the measure as specified
- The SMP reviewers agreed the validity tests and exclusions were appropriate, and results show high validity and clinical meaningfulness. Although, one member expressed concern regard finding replacing missing stroke severity scale with zeros
- The reviewers in general agreed that the risk adjustment analysis was thorough but one made a suggestion that an analysis of how the inclusion of socio-demographic risk factors affects the risk adjusted performance of these extreme providers would be more meaningful than simple differences in c-statistics for the overall population

Measure #2158: Medicare Spending Per Beneficiary (MSPB) - Hospital

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The MSPB Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital. Specifically, the MSPB Hospital measure assesses the cost to Medicare for Part A and Part B services performed by hospitals and other healthcare providers during an MSPB Hospital episode, which is comprised of the periods three days prior to, during, and 30 days following a patient's hospital stay. The MSPB Hospital measure is not condition specific and uses standardized prices when measuring costs. Beneficiary populations eligible for the MSPB Hospital calculation include Medicare beneficiaries enrolled in Medicare Parts A and B who were discharged between January 1 and December 1 in a calendar year from short-term acute hospitals paid under the Inpatient Prospective Payment System (IPPS).
- **Type of measure:** Cost/Resource Use
- **Data source:** Claims, Other
- **Level of analysis:** Facility
- **Risk-adjusted:** 109 risk factors adjusted and stratified by 26 risk categories; SES factors not included
- **Not based on a sample**
    - **Ratings for reliability:** H-7; M-0; L-0; I-0 → Measure passes with HIGH rating
    - Reliability testing conducted at the measure score level:
        - Signal-to-noise and multi-sample (or split-sample) analyses were conducted to assess reliability of the measure
        - The median reliability score for hospitals with at least 25 episodes was 0.96 and the reliability score interquartile range spanned from 0.91 to 0.98
        - The Pearson correlation coefficient was 0.83 for the 2018 split-sample and 0.79 for the 2017 and 2018 sample. The Shrout-Fleiss intraclass correlation coefficients were similar at 0.83 and 0.79 for the 2018 split-sample and 2017 and 2018 sample

        - The SMP did not raise any major concerns for reliability
- **Ratings for validity:** H-1; M-6; L-0; I-0 → Measure passes with MODERATE rating
- Validity testing conducted at the performance score level:
    - The developer assessed face validity via an expert panel and a review of public comments
    - Empirical testing consisted of comparing costs of episodes with and without post-acute services expected to increase cost
    - The developer reported the mean, standard deviation, and percentile distribution of observed to expected episode cost ratios for episodes with high-cost post-admission events higher than their counterparts as expected. "For example, episodes with an acute care rehospitalization an average O/E ratio of 1.55 and an interquartile range of 1.07 to 1.85, while episodes without such readmissions had an average O/E ratio of 0.89 and an interquartile range of 0.60 to 1.02."
    - Most service use/setting categories were moderately and positively correlated to the average predicted episode cost, with the correlations across all services categories average +0.487 and procedure use evidencing the strongest correlation +0.721. All three Payment & Value of Care measures, capturing 30-day Medicare payments for acute myocardial infarction, heart failure, and pneumonia conditions, were positively but weakly correlated with the hospital average predicted episode cost."
        - Some SMP concern of using standardized prices as a measure of resource use
    - Comparison of measure to other cost-specific and efficiency measures and measures in other HVBP program domains
    - Threats to validity:
        - For exclusions: Some concern regarding the roughly 37% of all episodes were excluded, with the largest contributor being episodes where the initial inpatient stay was in a non-acute hospital or a critical access hospital. Also, there were some concern about excluding patients who died.
        - No concerns with the ability to detect meaningful differences
        - For risk adjustment, the risk adjustment model followed the CMS Hierarchical Condition Category (HCC) risk adjustment methodology used in Medicare Advantage, including 79 HCC risk factors derived from claims 90 days prior to episode start date.
            - One SMP member mentioned that this is "somewhat concerning as 90 days is not typically sufficient time to see all patient's chronic conditions documented."
        - Social risk factors are included and comprise dual eligibility, race, sex, and SES from income, education and employment status and zip code

Subgroup 2

Measure #0229: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following heart failure (HF) hospitalization

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level 30-day, all-cause, risk-standardized mortality rate for patients discharged from the hospital with a principal diagnosis of HF. Mortality is defined as death for any cause within 30 days after the date of admission for the index admission. CMS annually reports the measure for patients who are 65 years or older and enrolled in FFS Medicare and hospitalized in non-federal hospitals or are patients hospitalized in VA facilities.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk-adjusted:** 24 risk factors; SES risk factors were tested but not included
- **Not based on a sample**
- **Ratings for reliability:** H-4; M-4; L-0; I-0 → Measure passes with MODERATE rating
    - Reliability testing conducted at the performance measure score level using both split-sample and signal-to-noise analysis (Adams' method)
    - Split-sample reliability score is 0.632
    - The median signal-to-noise reliability score of 0.79, ranging from 0.34 to 0.99
    - Regarding measure specifications, a SMP member asked for a clarification on the numerator: "Does this include both patients who were discharged alive and patients who were discharged dead?"
    - Regarding reliability testing: Most members agreed that the split-sample and signal-to-noise tests at the performance score level were appropriate, and the results show acceptable reliability
    - One reviewer voiced concerns about low reliability for the bottom 10% hospitals in the signal-to-noise ratio analysis ($r<0.44$), and split-sample reliability of 0.63 being perhaps acceptable but certainly not ideal. Question for the developer: "Can the developers comment on the impact of the observed reliability on misclassification or other consequences?"

- **Ratings for validity:** H-0; M-6; L-1; I-1 → Measure passes with MODERATE rating
    - Validity testing conducted at the performance measure score level, including both empirical validity testing (by comparing to CMS star ratings mortality scores, and star rating summary scores) and systematic assessment of face validity
    - The correlation between HF RSMRs and the Star-Rating mortality score is -0.676, which suggests that hospitals with lower HF RSMRs are more likely to have higher Star-Rating mortality scores
    - The correlation between HF RSMRs and Star-Rating summary score is -0.114, which suggests that hospitals with lower HF RSMRs are more likely to have higher Star-Rating summary scores
    - 24 clinical and demographic risk factors are included in the model; dual eligibility and AHRQ SES index were tested but not included in the final model
    - Members voiced no concerns about validity and noted the exclusions are appropriate

- However, a few members expressed concerns with the approach of demonstrating validity by using a comparator measure that includes the measure being tested; they suggested a stronger choice would be measures that do not already include the measure under study
- Some members pointed out the variation and low end of the distribution: 21% (992/4637) of hospitals had fewer than 25 cases and therefore could not be reliably assessed for their RSMR (risk-standardized mortality rate). Can developers elaborate on how the 25-case threshold was established?
- A couple of members asked for clarification on how model calibration is kept up to date using newer datasets
- A few members had questions about not including risk factors in the model because no added predictive power and no change in hospital performance rankings. It would be useful to know the rate of hospitals that would have change rank if social-risk factors would have been included, and the rationale behind why including other risk factors with nonsignificant coefficients did not apply to social risk factors

Measure #0531: Patient Safety Indicator (PSI) 90: Patient Safety and Adverse Events Composite

MEASURE HIGHLIGHTS Maintenance Measure

- **Description:** The PSI 90 composite measure summarizes patient safety across multiple indicators for the CMS Medicare fee-for-service population.
- **Type of measure:** Composite
- **Data source:** Claims
- **Level of analysis:** Facility
- **Risk-adjusted:** risk adjusted; social risk factors are not included
- **Not based on a sample**
- **Ratings for reliability:** H-2; M-5; L-0; I-1 → Measure passes MODERATE rating
- Reliability testing conducted with hospital-level data at the performance score level
    - For component reliability, ICCs were calculated, described as a signal-to-noise analysis. In addition, noise variance was computed for each component measure
        - The signal-to-noise reliability weight mean ranged from 0.167 on the low end for PSI 14 to 0.777 for PSI 03 (using the CMS v10.0)
        - Median reliability scores were somewhat lower with a low of 0.026 for PSI 14 to 0.668 for PSI 03
    - For composite reliability, split-sample testing was conducted and test-retest consistency.
        - The ICC for the composite was 0.74 (split-sample)
        - The ICC for the composite was 0.61 (test-retest)
    - Most analyses were performed with Medicare data from July 2016 to June 2019; confirmatory testing was performed with selected data from the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID)
- **Ratings for validity:** H-2; M-4; L-1; I-1 → Measure passes with MODERATE rating
- Validity testing conducted at the performance measure score level. There was also validity testing for each of the component measures and a face validity assessment.

  - In a 2019 TEP, members voted 12-1 in favor of continued use of PSI-90
  - Both component validity and composite validity testing were conducted. Component validity was performed to estimate the marginal effect of each PSI 90 component on harm in the Medicare population. Composite validity was assessed using convergent validity correlation with measure of patient safety published on the CMS website, as well as Leapfrog's Hospital Safety Survey. In addition, construct validity was conducted based on assessing how the measure correlated with the hospital resident-to-bed ratio, hospital nurse-to-bed ratio, and hospital nurse skill mix
    - For component validity, there was a significant impact of each of the PSI90 components on several adverse outcomes
    - For composite validity, there were significant correlations between PSI-90 and several Hospital Compare Measures that were in the expected direction
    - For convergent validity, there were significant correlations between PSI-90 and several readmission measures in the expected direction. There was little relationship between PSI 90 values and safe practice scores
    - There was no correlation between resident-FTE, nurse-FITE and nurse skill mix
- **Overall rating for the composite**: H-2; M-4; L-1; I-1 → Measure passes with MODERATE rating

Measure #0468: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Pneumonia Hospitalization

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level 30-day RSMR. Mortality is defined as death for any cause within 30 days after the date of admission for the index admission and discharged from the hospital with a principal discharge diagnosis of pneumonia, including aspiration pneumonia or a principal discharge diagnosis of sepsis (not severe sepsis) with a secondary diagnosis of pneumonia (including aspiration pneumonia) coded as POA. CMS annually reports the measure for patients who are 65 years or older and are either Medicare FFS beneficiaries and hospitalized in non-federal hospitals or patients hospitalized in VA facilities.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk-adjustment:** 36 risk factors adjusted; social risk factors were tested but not included in the final specification
- **Not based on a sample**
- **Ratings for reliability:** H-4; M-4; L-0; I-0 → Measure passes with HIGH rating
- Reliability testing conducted with hospital-level data at the performance measure score level:
  - Two types of reliability testing were conducted. First, an overall measure score reliability was calculated with the ICC using a split sample (i.e. test-retest) method. In addition, the facility-level reliability (signal-to-noise reliability) was also conducted
  - The data used for reliability testing was Medicare data (2016-2019), the American Community Survey (2013-2017), and the Master Beneficiary Summary File
    - The ICC for agreement between two independent assessments using the split-sample reliability was 0.668 (substantial agreement)

        - The mean reliability score was 0.78, ranging from 0.31 to 0.98 in the signal-to-noise reliability testing (substantial agreement)
- **Ratings for validity:** H-1; M-5; L-1; I-1 → Measure passes with MODERATE rating
- Validity testing conducted at the performance score level:
    - Validity testing conducted at the performance-measure score using both empirical validity testing and a systematic assessment of the face validity
    - Empirical validity was conducted comparing to two other measures: 1) the Hospital Star Rating mortality group score and the Overall Hospital Star Rating
        - In comparison to Star-Rating mortality scores, the correlation between the RSMR was -0.653
        - In comparison to the Overall Star-Rating scores, the correlation was -0.306.
    - It was stated that face validity was conducted with evaluation from a Technical Expert panel. No results were provided

Measure #1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level RSCR associated with elective primary THA and TKA in Medicare Fee-For-Service beneficiaries who are 65 years and older. The outcome (complication) is defined as any one of the specified complications occurring from the date of index admission to 90 days post date of the index admission (the admission included in the measure cohort). CMS annually reports the measure for patients who are 65 years or older, are enrolled in FFS Medicare and are hospitalized in non-federal acute-care hospitals.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data
- **Level of analysis:** Facility
- **Risk-adjustment:** 33 risk factors adjusted; social risk factors were tested but not included in the final specification
- **Not based on a sample**
    - **Ratings for reliability:** H-2; M-6; L-0; I-0 → Measure passes with MODERATE rating
    - Reliability testing conducted at the measure score level using an ICC on a split sample and by conducting a signal-to-noise analysis (Adams' method); minimum 25 procedures:
        - Signal-to-noise: 0.87 (median), 0.83 (mean), range (0.46 to 1.00). The 25th and 75th percentiles were 0.74 and 0.94, respectively
        - ICC: Using Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSCR for each hospital was 0.524. The developer states that this is a lower bound
        - The developer concluded substantial reliability for the measure score based on these results
        - Reviewers initially noted a concern around lack of clarity regarding the age range for the measure. The developer clarified that the measure is specified for the Medicare FFS 65+ population. Reviewers generally agreed that the testing approach and results were acceptable
    - **Ratings for validity:** H-0; M-6; L-1; I-1 → Measure passes with MODERATE rating

- Validity testing conducted at the measure score level. The measure was compared to Overall Hospital Star Rating and Hospital THA/TKA Surgical Volume:
    - The correlation between THA/TKA complications and Star-Rating summary score is -0.185, which suggests that hospitals with lower THA/TKA RSCRs are more likely to have higher Star-Rating summary scores especially at the extremes.
    - There is a general trend that high-volume hospitals (those in the upper deciles) have lower RSCRs than hospitals in other volume deciles.
    - Developer states, overall, the results above show that the trend and direction of this association is in line with what would be expected. Risk model discrimination and calibration: c statistic = 0.65; Developer reports good discrimination and predictive ability based on risk decile plot.
    - Reviewers generally accepted the validity testing results as a weak, but acceptable, demonstration of validity.

Measure #1551: Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level RSRR following elective primary THA and/or TKA in Medicare FFS beneficiaries who are 65 years and older. The outcome (readmission) is defined as unplanned readmission for any cause within 30 days of the discharge date for the index admission (the admission included in the measure cohort). A specified set of planned readmissions do not count in the readmission outcome. CMS annually reports the measure for patients who are 65 years or older, are enrolled in Medicare FFS, and hospitalized in non-federal acute-care hospitals.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data
- **Level of analysis:** Facility
- **Risk-adjustment:** 33 risk factors adjusted; social risk factors were tested but not included in the final specification
    - **Not based on a sample**
    - **Ratings for reliability:** H-2; M-5; L-1; I-0 → Measure passes with MODERATE rating
    - Reliability testing conducted at the measure score level using an ICC on a split sample and by conducting a signal-to-noise analysis (Adams' method); minimum 25 procedures:
        - Signal-to-noise: 0.77 (median), 0.72 (mean), range (0.29 to 0.99). The 25th and 75th percentiles were 0.58 and 0.88, respectively
        - ICC: Using Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSRR for each hospital was 0.454. Developer states this is a lower bound
        - Developer concludes substantial reliability for measure score based on results
        - Reviewers initially noted a concern around lack of clarity regarding the age range for the measure. The developer clarified that the measure is

specified for the Medicare FFS 65+ population. Reviewers generally agreed the testing approach and results were acceptable

- **Ratings for validity:** H-0; M-7; L-0; I-1 → Measure passes with MODERATE rating
- Validity testing conducted at the measure score level. The measure was compared to the Hospital Star Rating readmission group score, the Overall Hospital Star Rating, and Hospital THA/TKA Surgical Volume:
    - The correlation between THA/TKA RSRRs and Star-Rating readmissions score is -0.301, which suggests that hospitals with lower THA/TKA RSRRs are more likely to have higher Star-Rating readmission scores
    - The correlation between THA/TKA RSRRs and Star-Rating summary score is -0.239, which suggests that hospitals with lower THA/TKA RSRRs are more likely to have higher Star-Rating summary scores
    - Comparing various categories and deciles of hospital THA/TKA admission volume with THA/TKA readmission measure scores demonstrates an observed trend of higher hospital volume with lower readmission rates, especially in hospitals with the largest volumes
    - Developer states overall, the results above show that the trend and direction of these associations is in line with what would be expected
    - Risk model discrimination and calibration: c statistic = 0.67; Developer reports fair discrimination and predictive ability based on risk decile plot
    - Reviewers generally accepted the validity testing results as a weak, but acceptable, demonstration of validity.

Subgroup 3

Measure #3568: Person-Centered Primary Care Measure

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This measure is a patient-reported outcome (PRO) performance measure (PM). It is calculated using patient reported data gathered with the Person-Centered Primary Care Measure PCPCM PRO instrument. The PCPCM PRO is calculated as a continuous variable, based on a scale of one to four, in which a higher value equates to better quality. Responses to the PCPCM PRO instrument are aggregated at either the clinician or practice level to generate the PCPCM PRO-PM at that level. This process includes converting the scale to a zero to 100 point performance scale, in which a higher value continues to equate to better quality.
- **Type of measure:** Outcome: PRO-PM
- **Data source:** Instrument-Based Data
- **Level of analysis:** Clinician: Group/Practice, Clinician: Individual
- **Not risk-adjusted**
- **Sampling allowed**: The PCPCM PRO-PM performance score is calculated on all PCPCM PRO instruments received (eight of 11 items must have a response).
    - **Ratings for reliability:** H-2; M-3; L-1; I-2 →Measure passes MODERATE rating
    - Reliability testing conducted at the data element level:
        - Data element level testing was conducted using exploratory factor analysis, Rasch item fit statistics, and Cronbach's alpha testing

        - Exploratory Factor Analysis: This analysis resulted in the identification of a single factor: person centered primary care. This was further confirmed through calculated Eigenvalues of 6.9 for the patient online group and 4.7 for the patient point of care group.
        - Rasch Item Fit
            - Rasch item fit statistics ranged from 0.62 to 1.44 for the patient online group. Rasch item reliability was 0.99 for this group.
            - Rasch item fit statistics ranged from 0.55 to 1.49 for the patient point of care group. Rasch item reliability was 0.98 for this group.
        - Cronbach's alpha 0.91 for point of care group, 0.95 for the online group
    - Score level testing was conducted using ICC analysis between providers
        - Clinician: individual—ICC ranging between 0.76 and 0.94 and Guttman reliability ranging between 0.79 and 0.94
        - Group/practice—Split half reliability ranges from 0.86 and 0.95
- **Ratings for validity:** H-0; M-6; L-0; I-2 → Measure passes MODERATE rating
- Validity testing conducted at the data element level:
    - Data element testing was conducted using face validity for patients (n=30) who found the results to be meaningful to them and easy to answer. Clinicians (n=285) also found the results to be meaningful.
    - Data elements were further tested against two external measures of quality at the patient level: the Patient Enablement Instrument (PEI) and the What Matters Index (WMI).
    - Score level testing was conducted by comparing the PCPCM with
        - Clinician: individual—Pearson correlations between the PEI and the PCPCM at the clinician level ranging from 0.39 to 0.65
        - Group/practice—Pearson correlations between the PEI and the PCPCM are consistent from practice to practice, in the upper 0.5, and in agreement with the correlation for the total sample as well.
    - Threats to validity:
        - "The testing form claims no exclusions, thus the exclusion related questions in 2b2 are skipped. However, the MIF notes that completed instruments with less than 8 of the items completed are excluded. It would be been appropriate to state that exclusion in the testing form & respond to the corresponding questions regarding exclusions."
        - "The F-tests of homogeneity among practice scores and clinician scores are insufficient evidence of an ability to detect meaningful differences among units of analysis."
        - "Data are very favorably skewed with very small differences between highest and lowest performers for the 6 practices and 16 individual physicians for which data were reported. It is not clear if the differences that were considered as meaningful

(0.12 points different based on Cohen's d of 0.20) are actually meaningful as they are not calibrated against another validity variable. The paragraph describing statistical power is not sufficiently detailed to evaluate."

Measure #3567: Hemodialysis Vascular Access: Practitioner Level Long-term Catheter Rate Measure Title

MEASURE HIGHLIGHTS

- New Measure
- **Description:** Percentage of adult hemodialysis patient-months using a catheter continuously for three months or longer for vascular access attributable to an individual practitioner or group practice.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Clinician: Group/Practice, Clinician: Individual
- **Not risk-adjusted**
- **Not based on a sample**
    - **Ratings for reliability:** H-1; M-7; L-0; I-0 Pass→ Measure passes MODERATE rating
    - Reliability testing conducted at the score level:
        - Score level reliability testing conducted using inter-unit reliability (IUR) analysis as well as profile IUR (PIUR)
            - The IUR at practitioner level is 0.602. The PIUR at the practitioner level is 0.804
            - The IUR at practitioner group level is 0.793. The PIUR at the practitioner group level is 0.815
    - **Ratings for validity:** H-1; M-5; L-1; I-1 → Measure passes MODERATE rating
    - Validity testing conducted at the score level:
        - Validity was assessed using the trend test to measure the association between practitioner level long-term catheter rates occurring in January-December 2016, and hospitalization and mortality in the following 12 months
        - Clinician: individual level
            - Mortality rates are 17.0, 18.4, and 20.8 (per 100 patient-years) for practitioners having long-term catheter rates falling into the lowest 10%, middle, and highest 10% categories respectively ($p<0.001$)
            - Percentages of patient hospitalization (all-cause) are 60.8%, 62.8% and 67.8% for practitioners having long-term catheter rates falling into the lowest 10%, middle, and highest 10% categories respectively ($p<0.001$)
        - Clinician: group/practice level
            - Mortality rates are 18.4, 18.3, and 21.3 (per 100 patient-years) for practitioner-groups having long-term catheter rates falling into the lowest 10%, middle, and highest 10% categories respectively ($p<0.001$)

- Percentages of patient hospitalization (all cause) are 61.9%, 62.9% and 67.6% for practitioner-groups having long-term catheter rates falling into the lowest 10%, middle, and highest 10% categories respectively (p<0.001)

Measure #1623: Bereaved Family Survey

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure calculates the proportion of Veteran decedent's family members who rate overall satisfaction with the Veteran decedent's end-of-life care in an inpatient setting as "Excellent" versus "Very good", "Good", "Fair", or "Poor."
- **Type of measure:** Outcome: PRO-PM
- **Data source:** Instrument-Based Data
- **Level of analysis:** Facility, Other
- **Risk-adjusted:** The measure is risk-adjusted using five factors (veteran's age at the time of death; number of medical comorbidities present at the time of death; veteran's primary diagnosis on last admission; relationship of veteran's next-of-kin (i.e., spouse), and model of administration mode (i.e., mail). Social risk factors were not included in the final specification.
- **Sampling allowed**: A minimum sample size of 30 respondents is suggested to make comparisons between groups (facilities, VISNs).
    - **Ratings for reliability:** H-1; M-7; L-0; I-0 →Measure passes with MODERATE rating
    - Reliability testing conducted at the measure score and data element level:
        - To demonstrate reliability of the survey item used in this measure, the developers conducted four test-retest analyses on 93 randomly selected Bereaved Family Survey (BFS) respondents who agreed to complete the BFS on a second occasion (30 days apart)
            - Analysis #1 (Cohen's kappa): Kappa=0.5 (n=92); Developer cites Cohen's article that says a kappa of 0.5 indicates moderate agreement
            - Analysis #2: two-way random effects, absolute agreement, single rater/measurement: ICC (2,1) =0.52 (moderate agreement, according to Cohen)
            - Analysis #3, Logistic Regression: Compared to those who reported BFS=0 at time 1, respondents who reported BFS=1 at time 1 had 17.2 the odds of reporting BFS=1 at time 2 (interpreted as very strong association)
            - Analysis #4, Cohen's d Effect Size of a 2x2 contingency table: d=1.57 ("large" effect when d≥0.8)
        - Developers also described an analysis of the global item obtained via phone vs. mail administration (2009-2012 data for phone, 2012-2017 data for mail). Results indicate both are normally distributed (mean=58, 63 respectively, both with a standard deviation of 5), and very few facilities had mean ≥ 90, interpreted as no ceiling effects. Developers also reported Cronbach's alpha for phone vs. mail (0.81 vs 0.83).
        - To demonstrate reliability of the measure score, the developers conducted two analyses:

        - ICC1 using a mixed-effects logistic regression model:
            - FY10-FY12 (administered predominantly as a phone survey)
                - Facility-level variance estimate=0.15; 95% CI .12-.20; p<0.001
                - ICC1=0.04 (95% CI: .03-.06)
            - FY13-FY17 (administered predominantly via a mail survey):
                - Facility-level variance estimate =0.13; 95% CI .09-.20; p<0.001.
                - ICC1=0.04 (95% CI: .03-.06)
        - Split-half analysis with application of Spearman-Brown prophecy formula: 0.89
        - SMP members commented that reliability at the data element level is marginal and reliability at the measure score level is acceptable, but the reported ICC value of .04 is low. Because of the stronger measure score level testing findings, SMP members passed the measure on reliability.
- **Ratings for validity:** H-0; M-6; L-2; I-0 →Measure passes with MODERATE rating
- Validity testing conducted at the measure score and data element level:
    - To demonstrate validity of the survey item used in this measure, the developers analyzed five percent (randomly selected) of written responses to the question, "Is there anything else that you would like to share about the Veteran's care during the last month of life?" These comments were categorized as positive, neutral, or negative. These categorizations were correlated with the responses from the overall rating of care item (the item from the survey used in this measure).
        - Spearman correlation coefficient=0.51; p<0.001
    - Using patient-level data (N=84,616) and facility-level data (N=146), the developer ran nine separate logistic/linear regressions adjusted for nonresponse bias and patient case-mix. The independent variables were the process measures and the outcome variable was the individual BFS item and facility and patient level BFS percent "excellent." Their hypothesis was that receipt of each of the "best practices" processes should result in a statistically significant higher BFS score and the results of these analyses support the developer's hypotheses. Logistic regression analyses demonstrate statistically significant, positive associations between receipt of quality indicators, and patient-level BFS Performance Measure scores.
    - In the analysis of exclusions/missing data, a total of 16 percent of eligible decedent veterans were excluded from the measure. A total of 4 percent were excluded because they died within 24 hours of admission. The remaining excluded cases were included in a nonresponse bias analysis.
    - Prior to reporting of facility-level scores, the BFS-Performance Measure is adjusted for patient case-mix and survey nonresponse, and are stratified by facility complexity level. The measure is risk-adjusted using

five factors (veteran's age at the time of death; number of medical comorbidities present at the time of death; veteran's primary diagnosis on last admission; relationship of veteran's next-of-kin (i.e., spouse), and model of administration mode (i.e., mail).

Measure #3235: Hospice and Palliative Care Composite Process Measure—Comprehensive Assessment at Admission

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The Hospice Comprehensive Assessment Measure assesses the percentage of hospice stays in which patients who received a comprehensive patient assessment at hospice admission. The measure focuses on hospice patients age 18 years and older. A total of seven individual NQF endorsed component quality will provide the source data for this comprehensive assessment measure, including NQF #1634, NQF #1637, NQF #1639, NQF #1638, NQF #1617, NQF #1641, and NQF #1647. These seven measures are currently implemented in the CMS HQRP. These seven measures focus on care processes around hospice admission that are clinically recommended or required in the hospice Conditions of Participation, including patient preferences regarding life-sustaining treatments, care for spiritual and existential concerns, and management of pain, dyspnea, and bowels.
- **Type of measure:** Composite/process
- **Data source:** Other (Hospice Item Set)
- **Level of analysis:** Facility
- **Risk adjustment:** No risk adjustment; social risk factors were tested by not included in the final specification
- **Not based on a sample**
- **Ratings for reliability:** H-5; M-3; L-0; 0-I → Measure passes with HIGH rating
    - Reliability testing conducted at the performance score level
    - The ICC coefficient for this measure was 0.86
    - The signal-to-noise ratio for this measure was 3.55
    - Stability analysis showed that approximately 70% of providers had a change in QM score of less than one standard deviation
    - The SMP members generally agreed that the reliability tests are appropriate, and the results show moderate-to-high reliability, although two members raised questions about the signal-to-noise ratio method and how to interpret the resulting score of 3.55
- **Ratings for validity:** H-2; M-5; L-1; I-0 → Measure passes with MODERATE rating
    - Validity testing conducted at the individual performance measures level by examining correlations between each component measure and the overall composite measure
    - The p-values for all the Spearman correlation coefficients are significant (p-value < 0.01). There are significant positive correlations (ranging from 0.26 to 0.73) between the composite measure and each of the QMs
    - Exclusion analysis showed the mean QM scores among hospices were within 0.02 percentage points with and without the age exclusion
    - The missing rate for the majority of items were between 0.001 percent and 0.01 percent
    - The SMP reviewers generally agreed that this measure passed validity criterion; although, a couple pointed out it was coarse in its ability to identify "meaningful" differences in performance as the distribution is fairly compressed near the top
    - **Ratings for Composite:** H-2; M-6; L-0; I-0 → Measure passes with MODERATE rating

- The SMP reviewers generally agreed that the construction of the composite measure from the seven individual NQF-endorsed measures was straightforward

Measure #1893: Hospital 30-Day, all-cause, risk-standardized mortality rate (RSMR) following chronic obstructive pulmonary disease (COPD) hospitalization

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure estimates a hospital-level 30-day risk-standardized mortality rate (RSMR), defined as death from any cause within 30 days after the index admission date, for patients discharged from the hospital with either a principal discharge diagnosis of COPD or a principal discharge diagnosis of respiratory failure with a secondary discharge diagnosis of acute exacerbation of COPD. CMS annually reports the measure for patients who are 65 years or older and enrolled in FFS Medicare and hospitalized in non-federal hospitals or are patients hospitalized in VA facilities.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment Data, Other
- **Level of analysis:** Facility
- **Risk-adjustment:** 41 risk factors adjusted; social risk factors were tested but not included in the final specification
- **Not based on a sample**
    - **Ratings for reliability:** H-0; M-6; L-1; I-0 → Measure passes with MODERATE rating
        - Reliability testing conducted at the hospital level-data at the performance-score level. This was done by calculating CC using a split sample (i.e. test-retest) method
        - In addition, facility-level reliability (signal-to-noise reliability) was conducted
            - Split-Sample Reliability Agreement was 0.477 between two independent assessments (moderate agreement)
            - Signal-to-Noise, median reliability score was 0.72, range 0.32 to 0.97 (moderate reliability)
        - Data used for reliability testing was three years of Medicare administrative claims data (July 2016-June 2019). They also used the American Community Survey and the Master Beneficiary Summary File (MBSF)
    - **Ratings for validity:** H-2; M-5; L-0; I-0 → Measure passes with MODERATE rating
        - Validity testing conducted at the performance-measure score using both empirical validity testing and a systematic assessment of the face validity
        - Empirical validity was conducted comparing to two other measures: 1) the Hospital Star Rating mortality group score and the Overall Hospital Star Rating
            - In comparison to Star-Rating mortality scores, the correlation between the RSMR was -0.618
            - In comparison to the Overall Star-Rating scores, the correlation was -0.165

- Face validity was conducted with evaluation from a TEP. It appears this was the original face validity testing which was not updated
    - 90% of TEP members agreed or strongly agreed that measure was an accurate reflection of quality

Appendix B: Additional Information Submitted by Developers for Consideration

Subgroup 1

Measure #0330

Measure Title: Hospital 30-day all-cause risk-standardized readmission rate (RSRR) following heart failure (HF) hospitalization

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Split sample and signal-to-noise reliability**. A few Panel members disagreed with the developer's interpretation of the reliability testing results.
    - **Developer Response 1:** The split-sample reliability using two randomly split, non-overlapping samples, was 0.587. We interpret the split-sample reliability result as "moderate" based on the standards established by Landis and Koch (1977). This result is consistent with those from known, familiar, and related contexts. The current context is measuring provider quality, or specifically provider propensity to provide appropriate care as measured by subsequent outcomes. We identified several studies, which we think support the Landis & Koch guidelines when assessing test-retest reliability in the context of hospital measurement.
    - Hall et al calculated test-retest reliability for determining comorbidities from chart abstraction (Hall et al., 2006). In this study, multiple abstracters abstracted the same charts and the results were used to calculate four different common comorbidity scores. For three of the indices, test-retest reliabilities ranged from 0.59-0.68, with the fourth (the Charlson comorbidity score) achieving 0.80. We would argue that chart abstraction, with test-retest reliabilities in the 'moderate' to 'substantial' range, should be inherently more reliable than measuring hospital quality.
    - Cruz et al report reliabilities for collecting risk factor information from patients presenting to an emergency department with potential acute coronary syndrome (ACS) (Cruz et al., 2009). Each patient was queried twice, once by a clinician and once by a trained research assistant, and the reliabilities for a range of risk factors were calculated; these ranged from 0.28 (associated symptoms) to 0.69 (cardiac risk factors), with all other factors in the 0.30-0.56 range.
    - Hand et al report test-retest reliabilities for bedside clinical assessment of suspected stroke (Hand et al., 2006). Pairs of observers independently assessed suspected stroke

patients; findings were recorded on a standard form to promote consistency. The reliabilities were calculated for the full range of diagnostic factors: for vascular factors reliabilities ranged from 0.47-0.69 with only four of eight above 0.6; for history they ranged from 0.37-0.65 with only five of 12 above 0.6; other categories were similar (though reliability=1 for whether the patients were conscious).

Citations:

Cruz CO, Meshberg EB, Shofer FS, McCusker CM, Chang AM, Hollander JE. Interrater reliability and accuracy of clinicians and trained research assistants performing prospective data collection in emergency department patients with potential acute coronary syndrome. Ann Emerg Med. 2009 Jul;54(1):1-7.

Hall SF, Groome PA, Streiner DL, Rochon PA. Interrater reliability of measurements of comorbid illness should be reported. J Clin Epidemiol. 2006 Sep;59(9):926-33.

Hand PJ, Haisma JA, Kwan J, Lindley RI, Lamont B, Dennis MS, Wardlaw JM. Interobserver agreement for the bedside clinical assessment of suspected stroke. Stroke. 2006 Mar;37(3):776-80.

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

- **Issue 2: Variation/distribution of the measure score**. One panel member was concerned about the variation of the measure score and asked to see the distribution by deciles.
    - **Developer Response 2**: We submitted the distribution within the submission form, and we also show the distribution by decile below. As another reviewer noted, the range of the 10th to the 90th percentile is 3.4 percentage points, which is clinically meaningful. However, in addition to variation, a quality gap can be identified by overall poor performance, and an average readmission rate of 22% means that more than one out of every five patients with heart failure are returning to the hospital for an average performer, and for the worst performing hospitals, almost 1 out of every 3 patients are returning to the hospital within 30 days.

Distribution of Hospital Heart Failure RSRRs
Results for 07/2016-06/2019
Number of Hospitals: 4642
Number of Admissions: 1,286,352
Mean(SD)= 22(1.4)
Range(Min-Max): 16.7-31.2
Minimum: 16.7
10th percentile: 20.3
20th percentile: 21.0
30th percentile: 21.4
40th percentile: 21.6
50th percentile: 21.9
60th percentile: 22.1
70th percentile: 22.4
80th percentile: 22.9
90th percentile: 23.7
Maximum: 31.2

- **Issue 3: VA Data**. One Panel member asked about the VA population, and if the developer had any comparisons or testing results to see if the coefficients or models might be different in this population.
    - **Developer Response 3:** When VHA data were first added to the CMS readmission measures, extensive testing surrounding differences in cohorts, comorbidities and model fit were performed to determine if the addition of this unique population impacted the overall measure results. These analyses were shared with the VA and CMS, who determined that these measures could proceed with the VA population included for public reporting.
- **Issue 4: Insufficient testing for all-payer cohort**: Panel members noted that there are insufficient testing results to support the all-payer cohorts.
    - **Developer Response 4:** CORE has decided to change the measure specifications to limit the measure to the over 65, Medicare FFS patient population.

Validity

- **Issue 1: C-statistic:** One panel member noted that the c-statistic of 0.60 "may not be considered very good fit."
    - **Developer Response 1:** We ask that the Panel 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 comorbidities have a relatively limited relationship to the outcome 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. As noted by another SMP reviewer: "Readmission models tend to have lower C stats compared to mortality models because more of the mechanism for readmission lie outside the hospital." We also note that the risk-decile plots show that the model performance is acceptable.

- **Issue 2: Social risk factor adjustment interpretation of results:** One Panel member noted that the empiric data does not support the developer's decision not to include adjustment for social risk factors.
    - **Developer Response 2:** The decision to risk adjust is based both on the empiric results (impact on model and measure scores), the conceptual model (hospitals are better able to mitigate the influence of social risk factors on the measured outcome than clinicians) and CMS' policy decision to adjust for dual eligibility at the program level (within the Hospital Readmission Reduction Program or HRRP). In HRRP, hospitals are stratified into peer groups by the proportion of dual eligible patients and are penalized based on their performance within peer groups. Therefore adjustment for social risk factors affects payment to providers, but not the public reporting of the quality measures. It is also consistent with Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation's (ASPE's) recommendation that quality measures that are used for public reporting should not be risk adjusted (ASPE 2020). As noted in our submission, CMS also confidentially reports disparities in the readmission measures

to hospitals so that they have more detailed, actionable information about their patient population's social risk.

Citation:

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed October 2, 2020.

Measure #0505

Measure Title: Hospital 30-day all-cause risk-standardized readmission rate (RSRR) following acute myocardial infarction (AMI) hospitalization

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Reliability vote is consensus not reached**. Four Scientific Methods Panel members assessed the measure's reliability as "high" or "moderate," and four assessed the measure's reliability as "low."
    - **Developer Response 1:** The split-sample reliability using two randomly split, non-overlapping samples, was 0.424. We interpret the split-sample reliability result as "moderate" based on the standards established by Landis and Koch (1977). This result is consistent with those from known, familiar, and related contexts. The current context is measuring provider quality, or specifically provider propensity to provide appropriate care as measured by subsequent outcomes. We identified several studies, which we think support the Landis & Koch guidelines when assessing test-retest reliability in the context of hospital measurement.
    - Hall et al calculated test-retest reliability for determining comorbidities from chart abstraction (Hall et al., 2006). In this study, multiple abstracters abstracted the same charts and the results were used to calculate four different common comorbidity scores. For three of the indices, test-retest reliabilities ranged from 0.59-0.68, with the fourth (the Charlson comorbidity score) achieving 0.80. We would argue that chart abstraction, with test-retest reliabilities in the 'moderate' to 'substantial' range, should be inherently more reliable than measuring hospital quality.
    - Cruz et al report reliabilities for collecting risk factor information from patients presenting to an emergency department with potential acute coronary syndrome (ACS) (Cruz et al., 2009). Each patient was queried twice, once by a clinician and once by a trained research assistant, and the reliabilities for a range of risk factors were calculated; these ranged from 0.28 (associated symptoms) to 0.69 (cardiac risk factors), with all other factors in the 0.30-0.56 range.
    - Hand et al report test-retest reliabilities for bedside clinical assessment of suspected stroke (Hand et al., 2006). Pairs of observers independently assessed suspected stroke patients; findings were recorded on a standard form to promote consistency. The

reliabilities were calculated for the full range of diagnostic factors: for vascular factors reliabilities ranged from 0.47-0.69 with only four of eight above 0.6; for history they ranged from 0.37-0.65 with only five of 12 above 0.6; other categories were similar (though reliability=1 for whether the patients were conscious).

Citations:

Cruz CO, Meshberg EB, Shofer FS, McCusker CM, Chang AM, Hollander JE. Interrater reliability and accuracy of clinicians and trained research assistants performing prospective data collection in emergency department patients with potential acute coronary syndrome. Ann Emerg Med. 2009 Jul;54(1):1-7.

Hall SF, Groome PA, Streiner DL, Rochon PA. Interrater reliability of measurements of comorbid illness should be reported. J Clin Epidemiol. 2006 Sep;59(9):926-33.

Hand PJ, Haisma JA, Kwan J, Lindley RI, Lamont B, Dennis MS, Wardlaw JM. Interobserver agreement for the bedside clinical assessment of suspected stroke. Stroke. 2006 Mar;37(3):776-80.

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

- **Issue 2: Increasing the minimum case requirement.** A Panel member noted that reliability might be improved if the minimum number of cases was substantially raised.
    - **Developer Response 2**: CMS currently uses a 25-case cut off for all of the mortality and readmission measures. While we were not able to run additional reliability analyses with different case minimums within the response time for this feedback, this is an option we are interested in exploring further. Note that there is a tradeoff between raising the case count and reducing the number of eligible hospitals from public reporting.
- **Issue 3: VA data.** One Panel member asked about the VA population, and if the developer had any comparisons or testing results that would demonstrate if the coefficients or models might be different in this population.
    - **Developer Response 3:** When VHA data were first added to the CMS readmission measures, extensive testing surrounding differences in cohorts, comorbidities and model fit were performed to determine if the addition of this unique population impacted the overall measure results. These analyses were shared with the VA and CMS, who determined that these measures could proceed with the VA population included for public reporting.
- **Issue 4: Insufficient testing for all-payer cohort:** Reviewers noted that there are insufficient testing results to support the separate all-payer cohort.
    - **Developer Response 4**: CORE has decided to change the measure specifications to limit the measures to the over 65, Medicare FFS patient population.

Validity

- **Issue 1: Variation of the measure score:** One Panel member requested the distribution of measure scores in deciles.

- **Developer Response 1:** We provide the distribution in deciles below. Note that we provided this information in the submission form, but perhaps this was not included in the information provided to the Panel (as it is not within the "Intent to Submit" form). The median odds ratio for the measure score is 1.15, meaning that a patient has a 15% increase in the odds of a readmission at higher risk performance hospital compared to a lower risk hospital, suggesting meaningful variation in quality across hospitals.

Distribution of Hospital AMI RSRRs (All Hospitals)
Number of Hospitals: 4,074
Number of Admissions: 482,163
Mean(SD): 16.2(0.8)
Range(Min-Max): 11.5 - 22.9
Minimum: 11.5
10th percentile: 15.3
20th percentile:15.7
30th percentile: 15.9
40th percentile: 16.0
50th percentile: 16.1
60th percentile: 16.2
70th percentile: 16.3
80th percentile: 16.6
90th percentile: 17.1
Maximum: 22.9

- **Issue 2: Approach to social risk factor adjustment:** One Panel member questioned if the developer was following NQF guidance in the approach to social risk factor adjustment, in that "if clinical risk factors explain all or most of the patient variation in the outcome, then NQF guidance does not support adding social risk factors that do not account for variation."
  - **Developer Response 2:** In NQF's 2017 publication, they cite in "Recommendation 5" that like clinical risk factors, social risk factors should "contribute unique variation in the outcome" and should not be redundant. We note that this measure was initially endorsed in 2008 and the model was developed initially without social risk factors and the clinical risk factors that are currently in the model were included based on criteria used at the time of development.

    Citation:

    National Quality Forum. Evaluation of the NQF Trial Period for risk adjustment for social risk factors, 2017. Accessed October 2, 2020; Available at: https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=85635

- **Issue 3: Social risk factor adjustment interpretation of results:** One Panel member noted that the empiric data does not support the decision not to include adjustment for social risk factors.
  - **Developer Response 3:** The decision to risk adjust is based both on the empiric results (impact on model and measure scores), the conceptual model (hospitals are better able to mitigate the influence of social risk factors on the measured outcome than clinicians) and CMS' policy decision to adjust for dual eligibility at the program level (within the Hospital Readmission Reduction Program or HRRP). In HRRP, hospitals are stratified into peer groups by the proportion of dual eligible patients and are scored based on their

performance within peer groups. Therefore, adjustment for social risk factors affects payment to providers, but not the public reporting of the quality measures. It is also consistent with Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation's (ASPE's) recommendation that quality measures that are used for public reporting should not be risk adjusted (ASPE 2020). As noted in our submission, CMS also confidentially reports disparities in the readmission measures to hospitals so that they have more detailed, actionable information about their patient population's social risk.

Citation:

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed October 2, 2020.

- **Issue 4: C-statistic:** One panel member noted that the c-statistic "hovers around the mix-sixties, which makes one wonder overall goodness of fit of the model in terms of the included clinical and demographic factors."
    - **Developer Response 4:** We ask that the Panel 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 the outcome 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. As noted by another SMP reviewer: "Model discrimination [~0.65] is acceptable for a readmission model; readmission models tend to have lower C stats compared to mortality models because more of the mechanism for readmission lie outside the hospital."

- **Issue 5: Empiric validity.** One Panel member noted that CORE did not include the results of the chart-based model validation.
    - **Developer Response 5:** Thank you for noting that omission. Hospital-level adjusted readmission rates developed using the claims-based model were similar to rates produced for the same cohort using a medical record model (Krumholz, 2011). The slope of the weighted regression line between chart- and claims-based state readmission rates was 0.939 (SE, 0.0005), and the intercept was 0.011 (SE, 0.0001). The correlation coefficient of the standardized readmission rates from the 2 models was 0.98 (SE, 0.0006). The Spearman rank correlation coefficient was 0.9835. The median difference between the models in the hospital-specific risk-standardized readmission rates was 0.02 percentage points (25th percentile, −0.10; 75th percentile, 0.13; 10th percentile, −0.31; 90th percentile, 0.28).

      Citation:

      Krumholz HM, Lin Z, Drye EE, et al. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among

patients with acute myocardial infarction. *Circ Cardiovasc Qual Outcomes*. 2011;4(2):243-252.

- **Issue 6: Model validation.** Method panel members asked for additional information about how the developer validated its models with updated data.
    - **Developer Response 6:** Below we have provided more context related to our standard approach for model validation, and have addressed the specific comments from reviewers of other measures submitted by CORE that were shared with us via email. We use the same approach for validating models for all of the measures we maintain.
      Overall, please consider evaluating these measures in the context of re-endorsement; we are building on our original measure development work and provide additional information based on current data; the results we present in our testing attachment are about the original model and how it performs with current data. The strength of the evidence we submitted in the testing attachments (risk-decile plots and c-statistic) supports that the models remain valid for use with current data.
      CORE's measures undergo an annual measure reevaluation process, which ensures that the risk-standardized mortality models are continually assessed and remain valid, given possible changes in clinical practice and coding standards over time. Modifications made to measure cohorts, risk models, and outcomes are informed by review of the most recent literature related to measure conditions or outcomes, feedback from various stakeholders, and empirical analyses, including assessment of coding trends that reveal shifts in clinical practice or billing patterns. Input is solicited from a workgroup composed of up to 20 clinical and measure experts, inclusive of internal and external consultants and subcontractors.
      We provide a link to the 2020 measure re-evaluation report for this measure. The report describes what CORE did for 2020 public reporting; we:
        - Updated the ICD-10 code-based specifications used in the measures. Specifically, we:
        - Incorporated the code changes that occurred in the FY 2019 version of the ICD-10-CM/PCS (effective with October 1, 2018+ discharges) into the cohort definitions and risk models; and,
        - Applied a modified version of the FY 2019 V22 CMS-Hierarchical Condition Category (HCC) crosswalk that is maintained by RTI International to the risk models.
        - Monitored code frequencies to identify any warranted specification changes due to possible changes in coding practices and patterns;
        - Evaluated the stability of the risk-adjustment model over the three-year measurement period by examining the model variable frequencies, model coefficients, and the performance of the risk-adjustment model in each year (July 2016-June 2017, July 2017-June 2018, and July 2018-June 2019).
        - For each of the conditions, we assessed logistic regression model performance in terms of discriminant ability for each year of data and for the three-year combined period. We computed two summary statistics to assess model

performance: the predictive ability and the area under the receiver operating characteristic (ROC) curve (c-statistic).

In summary:

- We did not change which risk variables are in the model, but we updated the model coefficients using the updated dataset.
- We do compare the model coefficients over time; as described above, we calculate model coefficients for each year of the three-year period, and for the three years combined. See Table 4.2.2 in the measure re-evaluation report.
- Please note that risk-variable re-selection is part of our current workplan. To do this, we first needed a full three-year set of ICD-10-coded data, which we just reached this year. We had convened a Technical Expert Panel (TEP) for this work but COVID hit just as we were about to start. We delayed the project to allow the TEP (which includes clinicians) to attend to pressing COVID-related matters. In addition, due to COVID, CORE analytic staff have, and continue to have, restricted access to the physical building where these types of analyses must occur.

Other General Comments

- **Issue 1**: One Panel member asked us to clarify the number of hospitals with at least 25 cases.
  - **Developer Response 1**: Table 3 shows 2,161 hospitals, which represents the number of hospital with 25 or more admissions *before* applying the exclusion criteria because we are calculating the distribution of hospital percent of excluded cases. There are 1,932 hospitals with less than 25 cases *after* applying the exclusion criteria.

Measure #0506

Measure Title: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following pneumonia hospitalization

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Split sample and signal-to-noise reliability**. A few Panel members disagreed with the developer's interpretation of the reliability testing results.
  - **Developer Response 1:** The split-sample reliability using two randomly split, non-overlapping samples, was 0.544. We interpret the split-sample reliability result as "moderate" based on the standards established by Landis and Koch (1977). This result is consistent with those from known, familiar, and related contexts. The current context is measuring provider quality, or specifically provider propensity to provide appropriate care as measured by subsequent outcomes. We identified several studies, which we think support the Landis & Koch guidelines when assessing test-retest reliability in the context of hospital measurement.
  - Hall et al calculated test-retest reliability for determining comorbidities from chart abstraction (Hall et al., 2006). In this study, multiple abstracters abstracted the same

charts and the results were used to calculate four different common comorbidity scores. For three of the indices, test-retest reliabilities ranged from 0.59-0.68, with the fourth (the Charlson comorbidity score) achieving 0.80. We would argue that chart abstraction, with test-retest reliabilities in the 'moderate' to 'substantial' range, should be inherently more reliable than measuring hospital quality.

- Cruz et al report reliabilities for collecting risk factor information from patients presenting to an emergency department with potential acute coronary syndrome (ACS) (Cruz et al., 2009). Each patient was queried twice, once by a clinician and once by a trained research assistant, and the reliabilities for a range of risk factors were calculated; these ranged from 0.28 (associated symptoms) to 0.69 (cardiac risk factors), with all other factors in the 0.30-0.56 range.
- Hand et al report test-retest reliabilities for bedside clinical assessment of suspected stroke (Hand et al., 2006). Pairs of observers independently assessed suspected stroke patients; findings were recorded on a standard form to promote consistency. The reliabilities were calculated for the full range of diagnostic factors: for vascular factors reliabilities ranged from 0.47-0.69 with only four of eight above 0.6; for history they ranged from 0.37-0.65 with only five of 12 above 0.6; other categories were similar (though reliability=1 for whether the patients were conscious).

Citations:

Cruz CO, Meshberg EB, Shofer FS, McCusker CM, Chang AM, Hollander JE. Interrater reliability and accuracy of clinicians and trained research assistants performing prospective data collection in emergency department patients with potential acute coronary syndrome. Ann Emerg Med. 2009 Jul;54(1):1-7.

Hall SF, Groome PA, Streiner DL, Rochon PA. Interrater reliability of measurements of comorbid illness should be reported. J Clin Epidemiol. 2006 Sep;59(9):926-33.

Hand PJ, Haisma JA, Kwan J, Lindley RI, Lamont B, Dennis MS, Wardlaw JM. Interobserver agreement for the bedside clinical assessment of suspected stroke. Stroke. 2006 Mar;37(3):776-80.

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

- **Issue 2: Insufficient testing for all-payer cohort**: Panel members noted that there are insufficient testing results to support the all-payer cohorts.
  - **Developer Response 2:** CORE has decided to change the measure specifications to limit the measure to the over 65, Medicare FFS patient population.

Validity

- **Issue 1: Variation/distribution of the measure score**. One panel member was concerned about the variation of the measure score and asked to see the distribution by deciles.
  - **Developer Response 1**: We submitted the distribution within the submission form, and we also show the distribution by decile below. As another SMP reviewer noted, the range of the 10th to the 90th percentile is 2.6 percentage points, which is clinically

meaningful. However, in addition to variation, a quality gap can be identified by overall poor performance, and an average readmission rate of 16.7% means that one out of every six patients with pneumonia are returning to the hospital for an average performer, and for the worst performing hospitals, almost 1 in 4 patients are returning to the hospital within 30 days.

Distribution of Hospital Pneumonia RSRRs
Performance period: 07-2016-06/2019
Number of Hospitals: 4,697
Number of Admissions: 374,891
Mean (SD): 16.7 (1.1)
Range (min. – max.): 13.1 – 24.3
Minimum: 13.1
10th percentile: 15.4
20th percentile: 15.9
30th percentile: 16.1
40th percentile: 16.4
50th percentile: 16.6
60th percentile: 16.8
70th percentile: 17.0
80th percentile: 17.4
90th percentile: 18.0
Maximum: 24.3

- **Issue 2: VA Data**. One Panel member asked about the VA population, and if the developer had any comparisons or testing results to see if the coefficients or models might be different in this population.
    - **Developer Response 2**: When VHA data were first added to the CMS readmission measures, extensive testing surrounding differences in cohorts, comorbidities and model fit were performed to determine if the addition of this unique population impacted the overall measure results. These analyses were shared with the VA and CMS, who determined that these measures could proceed with the VA population included for public reporting.
- **Issue 3: Approach to social risk factor adjustment:** One Panel member questioned if the developer was following NQF guidance in the approach to social risk factor adjustment, in that "if clinical risk factors explain all or most of the patient variation in the outcome, then NQF guidance does not support adding social risk factors that do not account for variation."
    - **Developer Response 3:** In NQF's 2017 publication, they cite in "Recommendation 5" that like clinical risk factors, social risk factors should "contribute unique variation in the outcome" and should not be redundant. We note that this measure was initially endorsed in 2008 and the model was developed initially without social risk factors and the clinical risk factors that are currently in the model were included based on criteria used at the time of development.

      Citation:

National Quality Forum. Evaluation of the NQF Trial Period for risk adjustment for social risk factors, 2017. Accessed October 2, 2020; Available at: https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=85635

- **Issue 4: Social risk factor adjustment interpretation of results:** One Panel member noted that the empiric data does not support the developer's decision not to include adjustment for social risk factors.
    - **Developer Response 4:** The decision to risk adjust is based both on the empiric results (impact on model and measure scores), the conceptual model (hospitals are better able to mitigate the influence of social risk factors on the measured outcome than clinicians), and CMS' policy decision to adjust for dual eligibility at the program level (within the Hospital Readmission Reduction Program or HRRP). In HRRP, hospitals are stratified into peer groups by the proportion of dual eligible patients and are scored based on their performance within peer groups. Therefore adjustment for social risk factors affects payment to providers, but not the public reporting of the quality measures. It is also consistent with the Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation's (ASPE's) recommendation that quality measures that are used for public reporting should not be risk adjusted (ASPE 2020). As noted in our submission, CMS also confidentially reports disparities in the readmission measures to hospitals so that they have more detailed, actionable information about their patient population's social risk.

      Citation:

      Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed October 2, 2020.
- **Issue 5**: **Validation against medical records**. One Panel member noted that while we noted that we had performed validation against medical records for this measure, we did not present the results.
    - **Developer Response 5**: For the original version of this measure we validated the administrative model with a medical-record based model. The claims-based measure produced results which were highly correlated with those produced through manual chart audit: the correlation coefficient of estimated state-specific standardized readmission rates from the administrative and medical record models was 0.96 (Lindenauer et al., 2011).

      Citation:

      Lindenauer PK, Normand SL, Drye EE, Lin Z, Goodrich K, Desai MM, Bratzler DW, O'Donnell WJ, Metersky ML, Krumholz HM. Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. J Hosp Med. 2011 Mar;6(3):142-50.

- **Issue 6: Model validation.** Method panel members asked for additional information about how the developer validated its models with updated data.

- **Developer Response 6:** Below we have provided more context related to our standard approach for model validation, and have addressed the specific comments from reviewers of other measures submitted by CORE that were shared with us via email. We use the same approach for validating models for all of the measures we maintain.
  Overall, please consider evaluating these measures in the context of re-endorsement; we are building on our original measure development work and provide additional information based on current data; the results we present in our testing attachment are about the original model and how it performs with current data. The strength of the evidence we submitted in the testing attachments (risk-decile plots and c-statistic) supports that the models remain valid for use with current data.
  CORE's measures undergo an annual measure reevaluation process, which ensures that the risk-standardized mortality models are continually assessed and remain valid, given possible changes in clinical practice and coding standards over time. Modifications made to measure cohorts, risk models, and outcomes are informed by review of the most recent literature related to measure conditions or outcomes, feedback from various stakeholders, and empirical analyses, including assessment of coding trends that reveal shifts in clinical practice or billing patterns. Input is solicited from a workgroup composed of up to 20 clinical and measure experts, inclusive of internal and external consultants and subcontractors.
  We provide a link to the 2020 measure re-evaluation report for this measure. The report describes what CORE did for 2020 public reporting; we:
  - Updated the ICD-10 code-based specifications used in the measures. Specifically, we:
  - Incorporated the code changes that occurred in the FY 2019 version of the ICD-10-CM/PCS (effective with October 1, 2018+ discharges) into the cohort definitions and risk models; and,
  - Applied a modified version of the FY 2019 V22 CMS-Hierarchical Condition Category (HCC) crosswalk that is maintained by RTI International to the risk models.
  - Monitored code frequencies to identify any warranted specification changes due to possible changes in coding practices and patterns;
  - Evaluated the stability of the risk-adjustment model over the three-year measurement period by examining the model variable frequencies, model coefficients, and the performance of the risk-adjustment model in each year (July 2016-June 2017, July 2017-June 2018, and July 2018-June 2019).
  - For each of the conditions, we assessed logistic regression model performance in terms of discriminant ability for each year of data and for the three-year combined period. We computed two summary statistics to assess model performance: the predictive ability and the area under the receiver operating characteristic (ROC) curve (c-statistic).

  In summary:
  - We did not change which risk variables are in the model, but we updated the model coefficients using the updated dataset.

- We do compare the model coefficients over time; as described above, we calculate model coefficients for each year of the three-year period, and for the three years combined. See Table 4.5.2 in the measure re-evaluation report.
- Please note that risk-variable re-selection is part of our current workplan. To do this, we first needed a full three-year set of ICD-10-coded data, which we just reached this year. We had convened a Technical Expert Panel (TEP) for this work but COVID hit just as we were about to start. We delayed the project to allow the TEP (which includes clinicians) to attend to pressing COVID-related matters. In addition, due to COVID, CORE analytic staff have, and continue to have, restricted access to the physical building where these types of analyses must occur.

Measure #3597

Measure Title: Clinician-Group Risk-Standardized Acute Hospital Admission Rate for Patients with Multiple Chronic Conditions under the Merit-based Incentive Payment System

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Specifications:** One Panel member asked if, based on the title, this measure will only be used for the Merit-based Incentive Payment System (MIPS) or if the intent was to use the measure more widely. The reviewer states: "if the latter is the case, then these words should be deleted from the measure title. If the former is the case, then the Denominator state should make explicit that only those Providers should be included in the measure calculation.**"**
    - **Developer Response 1:** Yes, this measure is intended to be used in the Merit-based Incentive Payment System, and is specified (and was developed and tested) for use for MIPS-eligible providers. Patients not attributed to a MIPS-eligible provider are removed from the measure as stated in the denominator exclusions.

Validity

- **Issue 1: Meaningful Differences:** One Panel member asked for clarification on how the median rate ratio or MRR was calculated and how this is done without any specific thresholds for higher- and lower-risk providers.
    - **Developer Response 1:** The median rate ratio is the median relative change in the rate of the occurrence of an event (in this case, an admission) when comparing the same patient attributed to 2 randomly selected providers in different clusters that are ordered by their measure score (Austin, 2018). Therefore, no specific threshold is needed for defining the two clusters.

      Citation

      Austin, PC, Stryhn, H, Leckie, G, Merlo, J. Measures of clustering and heterogeneity in multilevel Poisson regression analyses of rates/count data. Statistics in Medicine. 2018; 37: 572– 589. https://doi.org/10.1002/sim.7532

- **Issue 2: Interpretation of the measure score:** One Panel member asked for clarification regarding a statement made by the developer that: "Across the 4,044 clinician groups who had at least one MCC patient, RSAAR measure scores, including adjustment for the social risk factors of AHRQ SES Index, and physician-specialist density, ranged from 20.4 to 98.7 per 100 person-years, with a median of 40.4 and an IQR of 36.0 to 45.2. This indicates that after adjustment half of Medicare patients with multiple chronic conditions had between 36 and 45 acute care visits in a year." The reviewer asked if the developer meant per 100 person years rather than "in a year."
    - **Developer Response 2:** Thank you for pointing this out. You are correct that we should have said per "100-person years" rather than "in a year"
- **Issue 3**: **Social risk factor adjustment**: One reviewer noted that the developer's results show that SES should be a risk factor in the model and another reviewer noted that the rationale for risk adjusting the MIPS MCC measure was in conflict with the rationale presented for CORE's hospital-level readmission measures going through NQF endorsement. [Note that in this MIPS MCC response we provide a discussion below that addresses the MIPS MCC, ACO MCC and hospital-level readmission measures.]
    - **Developer response 3**: We thank the reviewer for their comments. Please note that CMS did decide to include two social risk factors (low physician-specialist density, and low AHRQ SES index) in both the MIPS and ACO MCC measures, however the inclusion of these two social risk factors in the ACO MCC updated measure was done primarily to fully align the measure with the MIPS clinician-group level MCC measure. As discussed in the MIPS MCC and ACO measure NQF applications, CMS took a wholistic look, both quantitative and qualitative, at the benefits and risks of adjusting for social risk factors in each of the measures and balances sometimes competing considerations in reaching a decision on whether to include social risk factor adjustment in the models. CMS included the AHRQ SES index and physician-specialist density variables in the MIPS MCC measure because MIPS-eligible clinicians working in the community may have a limited ability to influence these community-based contextual factors that affect admission risk. However, CMS acknowledges that ACOs typically have a greater ability to mitigate the increased admission risk associated with social risk factors than smaller clinician groups that will be measured by the MIPS MCC measure. On balance, however, CMS decided it was programmatically very important to fully align the two admission measures' specifications given their use in the two QPP programs and the possibility that individual providers and groups may be assessed by either or both measures over time given their financial arrangements with clinician groups and ACOs.

      The reviewer is right to point out that while CMS did decide to add low AHRQ SES index to the risk adjustment model for the clinician-group-level MCC measure, they did not adjust the hospital-level readmission measures. CMS decisions were informed by the conceptual models for each measure that specify the pathways through which social risk factors influence the measured outcome. CMS also considered the literature and results of empiric analyses.

      The difference in the approach to the hospital readmission and MIPS/ACO admission measures is driven largely by the different context in which the outcomes are being measured. The ACO and MIPS MCC measures assess clinicians' ability to manage

patients over the course of a year in the ambulatory setting in a way that minimizes acute unplanned admissions. The conceptual model for these measures acknowledges that social risk factors influence the outcome, and clinicians in the community may not be able to fully mitigate the pathways through which this influence occurs throughout the course of a year given their limited interactions with their patients. In contrast, the hospital readmission measures assess a near-term (30-day) outcome of an acute episode of care, during which generally well-resourced institutions have the patient under their care and can identify, plan for and mitigate specific factors that put patients at increased risk of readmission. In other words, hospitals have a greater ability to mitigate the influence of social risk factors on the measured outcome. In fact, quality improvement efforts to reduce 30-day hospital readmissions focus on those very areas that hospitals can control, such as improving care transitions, including patient education, medication reconciliation, discharge planning, and coordination of outpatient care, which have been shown to reduce readmission rates.

- **Issue 4: Model performance.** One Panel member noted that the variance R squared of 0.105 was not an indication of a very strong model.
    - **Developer Response 4:** Thank you for your comment. Please note that another SMP reviewer pointed out that while this is low, it is consistent with other risk adjustment models that have been endorsed by NQF.
- **Issue 5: Qualifying conditions and validity:** One reviewer asked why the cohort qualifying conditions include certain chronic conditions related to the nervous, endocrine, cardiovascular, respiratory, and digestive systems but exclude other chronic conditions related to the integumentary, skeletal, muscular, lymphatic, urinary, and reproductive systems.
    - **Developer response 5**: We targeted applied transparent criteria with input from stakeholders, building on the NQF MCC framework to define criteria for cohort-defining chronic conditions. In brief: (1) we considered a condition "chronic" if it was persistent and had an adverse effect on health status, function, or quality of life (and was not an asymptomatic risk factor); (2) necessitated complex health care management, decision-making, or coordination when co-occurring with other conditions; and (3) high-quality ambulatory ACO care could lower the risk of admission among patients with these chronic conditions. Moreover, we empirically tested combinations of chronic conditions with respect to their prevalence and associated risk of admission and published the approach in the medical literature (Drye et al., 2018). Patients with other chronic conditions, such as those related to the integumentary, skeletal, muscular, lymphatic, urinary, and reproductive systems are not excluded but these conditions are not used as cohort-defining conditions.

      Citation:

      Drye, E. E. , Altaf, F. K. , Lipska, K. J. , Spatz, E. S. , Montague, J. A. , Bao, H. , Parzynski, C. S. , Ross, J. S. , Bernheim, S. M. , Krumholz, H. M. & Lin, Z. (2018). Defining Multiple Chronic Conditions for Quality Measurement. Medical Care, 56(2), 193–201.
- **Issue 6: Measure Validity:** Two panel members commented on the fact that the developer provided face validity as evidence of validity. One reviewer noted that the developer could provide literature that demonstrates that increased admissions are a quality signal.

- **Developer response 6:** Note that for new measures, NQF permits face validity as evidence of measure validity. Also, typically developers provide supporting literature in the evidence attachment, which is not submitted to the Scientific Methods Panel for their review. We have provided this measure's evidence attachment that shows there is strong evidence supporting the assertion that ambulatory care clinicians can influence admission rates through quality of care.

Measure #3596

Measure Title: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1:** A panel member noted that the measure testing included hospitals that report the NIH Stroke Scale for at least 60% of ischemic stroke admissions and while that is a relatively low threshold, only 329 hospitals had enough admissions with the NIH Stroke Scale to be included.
    - **Developer Response 1:** This measure was re-specified from the existing stroke mortality measure within the Inpatient Quality Reporting (IQR) Program in response to stakeholder input that the measure should adjust for stroke severity upon admission to better reflect the hospital's ability to influence survival following acute ischemic stroke. Hospital reporting of the National Institutes of Health (NIH) Stroke Scale aligns with guidelines and recommendations put forth by the American Heart Association (AHA) and American Stroke Association (ASA). However, hospital reporting of the NIH Stroke Scale within International Classification of Diseases, Tenth Revision (ICD-10) claims has been low (13% in October 2016 when ICD-10 claims became available) but increasing (56% in Mary 2019). For testing, we used all 33 months of available ICD-10 data to mirror the intended measurement period of three years, requiring us to utilize older data with lower NIH Stroke Scale reporting. CMS anticipates reporting of the NIH Stroke Scale will continue to increase with a more sufficient rate of NIH Stroke Scale reporting by the time the measure is implemented within IQR for fiscal year 2023 payment determination, as signaled by CMS within the Fiscal Year (FY) 2018 Inpatient Prospective Payment System (IPPS) Rule. For all hospitals in the testing sample (N=329), the median signal to noise reliability score was 0.72 and for hospitals with at least 25 cases (N=292), the median signal to noise reliability score was 0.75.
- **Issue 2:** Data element reliability was indicated but panel members were unable to locate the results.
    - **Developer Response 2:** For section 2a2. RELIABILITY TESTING, the "performance measure score" box should have been selected instead of the "critical data elements used in the measure". While this was an error on our part while completing the forms, our understanding is that data element reliability is not required for endorsement. With that being said, the measure utilizes only data elements from administrative claims that are consequential for payment and consistently audited. CMS has in place several

hospital auditing mechanisms used to assess overall claims code accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures, including diagnosis and procedure codes and other elements that are consequential to payment.

- **Issue 3:** A panel member noted that we provided reliability for both all hospitals and hospitals with at least 25 cases and requested clarification on whether the measure is specified for hospitals with 0 or 25 cases. Under validity, another panel member wondered why we did not consider the minimum case count of 25 as an exclusion criterion before calculating measure scores and conducting measure testing.
    - **Developer Response 3:** Consistent with many of CMS's mortality and readmission outcome measures and given the comparative methodology, scores are calculated using data from all hospitals. However, measure scores would only be assigned to and publicly reported for hospitals with at least 25 cases. The minimum case count is not a measure exclusion; rather, a reporting threshold to ensure that only hospitals with sufficient information on ischemic stroke patients receive a measure score. As done for the current stroke mortality measure within IQR and other CMS mortality and readmission outcome measures, hospitals with fewer than 25 cases would still receive hospital specific reports with patient-level data and measure results for transparency and performance improvement.
- **Issue 4:** There was some concern about reliability in low volume hospitals, and that the results were not stratified by hospital volume.
    - **Developer Response 4:** Variation in volume can impact reliability. However, consistent with CMS's other mortality and readmission outcome measures, measure scores would only be assigned and publicly reported for hospitals with at least 25 cases. This ensures quality information be available for most hospitals while maintaining reliable measure scores. We used the formula presented by Adams and colleagues (2010) to calculate the facility-level reliability.[a] The median reliability score was 0.75, ranging from 0.24 to 0.95. The 25th and 75th percentiles were 0.59 and 0.83, respectively. The median reliability score demonstrates sufficient reliability. We also report confidence intervals for measure results that account for volume.

Validity

- **Issue 1:** Some panel members noted that the range of risk-standardized mortality rates (RSMR)s was fairly narrow and questioned whether the small difference was enough to reliably measure relative performance.
    - **Developer Response 1:** While the range of RSMRs is fairly narrow, with 10th-90th percentile RSMRs of 13.04-16.28, mortality is an important health outcome that is meaningful to patient and providers. Therefore, even small variations in mortality rates across hospitals indicate opportunities for improvement and potential lives saved. For

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

this measure, the median odds ratio is the median relative change in the rate of the occurrence of an event (in this case, an admission) when comparing the same patient attributed to two randomly selected providers in different clusters that are ordered by their measure scores. Therefore, no specific threshold is needed for defining the two clusters. The median odds ratio using the between hospital variance is 1.21, indicating the measure is capable of identifying meaningful differences in hospital performance.

- **Issue 2:** One panel member expressed "significant" concerns with using simple replacement with zero to address missing NIH Stroke Scale values. They inquired whether the comparison between the c-statistics for the stroke mortality measure currently within IQR and this measure with simple replacement with zero are valid. They also noted potential shifts in hospital performance, and therefore payment, should this measure and simple replacement with zero method replace the current stroke mortality measure within IQR.
    - **Developer Response 2:** Hospital reporting of the NIH Stroke Scale aligns with guidelines put forth by the AHA and ASA. However, reporting of the NIH Stroke Scale remains relatively low but gradually increasing. Simple replacement with zero would further incentivize reporting of the NIH Stroke Scale, as recommended by the AHA and ASA. CMS may use multiple imputation, as outlined within this measure's submission, once hospitals meet a sufficient threshold of NIH Stroke Scale reporting. The comparison of model c-statistics for the stroke mortality measure currently reported within IQR and this measure with simple replacement with zero was meant to demonstrate that, regardless of the approach to addressing missing NIH Stroke Scale values, including adjustment for stroke severity improves model performance. While shifts in hospital performance may be observed should this measure utilize simple replacement with zero and replace the stroke mortality measure currently reported within IQR, this measure was re-specified in response to extensive stakeholder input to adjust for stroke severity in order to better reflect hospitals' ability to influence survival.
- **Issue 3:** One panel member suggested it would have been helpful if we compared characteristics of hospitals who did and did not have report the NIH Stroke Scale for at least 60% of their ischemic stroke patients to reveal systematic biases.
    - **Developer Response 3:** Hospital reporting of the NIH Stroke Scale is a standard of care that aligns with guidelines put forth by the AHA and ASA. It is unclear why some hospitals do not report the NIH Stroke Scale within ICD 10 claims. However, reporting has increased from 13% in October 2016 when ICD-10 codes were first available to 56% in May 2019. We demonstrate this increase in NIH Stroke Scale reporting on the level of hospital characteristics by showing the frequency and percent of hospitals that report the NIH Stroke Scale for at least 60% of ischemic stroke patients in the testing sample (33 months; October 2016 – June 2019) and within the most recent year of data (July 2018 – June 2019). While fewer small, teaching, suburban and urban, safety-net, and critical access hospitals report the NIH Stroke Scale for at least 60% of ischemic stroke patients within the testing sample, there is an observed increase in NIH Stroke Scale reporting across all of the hospital characteristics within the most recent year of data. Given these increases, we expect that the proportions of hospitals reporting the stroke scale will continue to increase over time across all hospital characteristics.

- **Issue 4:** One panel member recommended an analysis of how the risk-adjusted performance of providers is affected at the low and high extremes of proportion of social risk factors.
    - **Developer Response 4:** Based on our analyses, we know that the median proportion of dual-eligible patients is 11.19% with a range of 0-100% (interquartile range [IQR] 7.69-16.85%) and low AHRQ SES is 12.03% with a range of 0-64.29% (IQR 5.26-23.08%). We also know that hospitals with the lowest quartile of dual-eligible patients had a median RSMR of 14.53% with a range of 11.72-17.21% (IQR 14.05-15.04%) and similarly, hospitals with the highest quartile of dual-eligible patients had a median RSMR of 14.33% with a range of 10.05-17.44% (IQR 13.90-14.82%). Hospitals in the lowest quartile of low AHRQ SES had a median RSMR of 14.52% with a range of 12.34-16.16% (IQR 14.14-14.82%) and similarly, hospitals in the highest quartile of low AHRQ SES had a median RSMR of 14.51% with a range of 11.76-17.83% (IQR 13.86-15.32%). These results demonstrate that there are similar rates of stroke mortality after adjustment for clinical risk factors, regardless of social risk factors. While we did not specifically isolate the performance after social risk factor adjustment for hospitals with extreme low and high proportions of patients with social risk factors, we did analyze impact of adding social risk factors on the risk-standardized measure scores. The mean absolute change in hospitals' RSMRs when adding a dual-eligibility indicator was 0.001% with a correlation coefficient between RSMRs for each hospital with and without dual-eligibility of 0.999. The mean absolute change in hospitals' RSMRs when adding a low SES AHRQ indicator was 0.00% with a correlation coefficient between RSMRs for each hospital with and without low SES of 0.999.
- **Issue 5:** Some panel members requested additional information about face validity testing, including work group composition and their potential conflicts of interest as well as how the advisory group was polled for face validity.
    - **Developer Response 5:** This measure was originally developed in 2016 in response to urgent feedback from stakeholders to adjust stroke mortality scores for stroke severity upon admission. The work group was convened and polled for face validity in 2016. The work group consisted of neurologists, cardiologists, and experts in biostatistics, measurement, and quality improvement, including Lee Schwamm, MD, who was Vice Charmain of the Department of Neurology at Massachusetts General Hospital, Gregg Fonarow, MD, who was Professor of Medicine at the University of California, Jason Sico, MD, who was Director of Stroke Care within the VA Connecticut Healthcare System, and Kevin Sheth, MD, who was Associate Professor of Neurology and Neurosurgery at Yale University. Please note that work group member titles and associations may have changed over time. The work group met regularly throughout development to address key issues related to measure cohort, outcome, and usability. In addition, in 2016, we also posted the measure specifications for public comment, which resulted in overall agreement on the face validity of the measure and the inclusion of the NIH Stroke Scale as a risk adjustment. While the public comment summary report is no longer available on CMS's website, we can circulate the report to panel members upon request.
- **Issue 6:** Some panel members questioned the use of the Overall Star Rating mortality measure group to assess measure score validity.

- **Developer Response 6:** We provided broad assessment of validity for this measure, including data element validity, measure score validity, and face validity. When researching options for measure scale validity, we found few external stroke mortality metrics with available national data that have been widely-regarded or demonstrated as valid. The Overall Star Ratings mortality measure group is meant to provide a summary of hospital performance on seven available mortality measures, including the stroke mortality measure currently reported within IQR. A moderate correlation is to be expected between the Overall Star Rating mortality measure group and this measure since they both assess mortality, but the Overall Star Ratings uses different cohorts, multiple metrics, and a complex statistical model to determine which measures should contribute most to the group score. In addition to measure score validity, we demonstrated excellent data element validity by comparing NIH stroke scale scores from claims to scores in the Get With The Guidelines (GWTG) registry. Although not presented within the submission for this measure, we did compare hospital RSMRs calculated using NIH Stroke Scale scores from GWTG registry and ICD-10 claims, resulting in a 0.922 correlation.

Other General Comments

- **Issue 1**: On page 1 of the SMP Preliminary Assessment, the checkbox for "previously endorsed" was selected.
  - **Developer Response 1:** This measure is being submitted for initial endorsement and has not been previously endorsed by NQF.

Measure #1891

Measure Title: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following chronic obstructive pulmonary disease (COPD) hospitalization

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Interpretation of split-sample and signal-to-noise reliability testing results**. A few Panel members disagreed with the developer's interpretation of the reliability testing results.
  - **Developer Response 1**: The split-sample reliability using two randomly split, non-overlapping samples, was 0.406. We interpret the split-sample reliability result as "moderate" based on the standards established by Landis and Koch (1977). This result is consistent with those from known, familiar, and related contexts. The current context is measuring provider quality, or specifically provider propensity to provide appropriate care as measured by subsequent outcomes. We identified several studies, which we think support the Landis & Koch guidelines when assessing test-retest reliability in the context of hospital measurement.
  - Hall et al calculated test-retest reliability for determining comorbidities from chart abstraction (Hall et al., 2006). In this study, multiple abstracters abstracted the same charts and the results were used to calculate four different common comorbidity scores. For three of the indices, test-retest reliabilities ranged from 0.59-0.68, with the fourth

(the Charlson comorbidity score) achieving 0.80. We would argue that chart abstraction, with test-retest reliabilities in the 'moderate' to 'substantial' range, should be inherently more reliable than measuring hospital quality.

- Cruz et al report reliabilities for collecting risk factor information from patients presenting to an emergency department with potential acute coronary syndrome (ACS) (Cruz et al., 2009). Each patient was queried twice, once by a clinician and once by a trained research assistant, and the reliabilities for a range of risk factors were calculated; these ranged from 0.28 (associated symptoms) to 0.69 (cardiac risk factors), with all other factors in the 0.30-0.56 range.
- Hand et al report test-retest reliabilities for bedside clinical assessment of suspected stroke (Hand et al., 2006). Pairs of observers independently assessed suspected stroke patients; findings were recorded on a standard form to promote consistency. The reliabilities were calculated for the full range of diagnostic factors: for vascular factors, reliabilities ranged from 0.47-0.69 with only four of eight above 0.6; for history they ranged from 0.37-0.65 with only five of 12 above 0.6; other categories were similar (though reliability = 1 for whether the patients were conscious).

Citations:

Cruz CO, Meshberg EB, Shofer FS, McCusker CM, Chang AM, Hollander JE. Interrater reliability and accuracy of clinicians and trained research assistants performing prospective data collection in emergency department patients with potential acute coronary syndrome. Ann Emerg Med. 2009 Jul;54(1):1-7.

Hall SF, Groome PA, Streiner DL, Rochon PA. Interrater reliability of measurements of comorbid illness should be reported. J Clin Epidemiol. 2006 Sep;59(9):926-33.

Hand PJ, Haisma JA, Kwan J, Lindley RI, Lamont B, Dennis MS, Wardlaw JM. Interobserver agreement for the bedside clinical assessment of suspected stroke. Stroke. 2006 Mar;37(3):776-80.

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

- **Issue 2: Increasing the minimum case requirement.** A Panel member noted that reliability might be improved if the minimum number of cases was substantially raised.
  - **Developer Response 2**: CMS currently uses a 25-case cut off for all of the mortality and readmission measures. While we were not able to run additional reliability analyses with different case minimums within the response time for this feedback, this is an option we are interested in exploring further. Note that there is a tradeoff between raising the case count and reducing the number of eligible hospitals from public reporting.
- **Issue 3: Insufficient testing for all-payer cohort.** Reviewers noted that there are insufficient testing results to support the separate all-payer cohort.
  - **Developer Response 3**: CORE has decided to change the measure specifications to limit the measures to the over 65, Medicare FFS patient population.

Validity

- **Issue 1: Low numbers of outliers identified by performance categories.** A Panel member expressed concern about the relatively low proportion of performance outliers. Another Panel member asked the developer to provide the distribution of measure scores.
    - **Developer Response 1:** Please note that performance categories are an implementation issue – CMS chooses to identify outliers based on 95% interval estimates, akin to 95% confidence intervals, which is a conservative approach to identifying performance outliers. These categories are also based on a comparison to the national average.

      Below we show the distribution of measure scores; a Panel member noted that the range of performance between the 10th and the 90th percentile is clinically meaningful. In addition, a quality gap can also be identified by overall poor performance: an average readmission rate of 19.6% means that almost one out of every five patients with COPD are returning to the hospital for an average performer, and for the worst performing hospitals, more than 1 in 4 patients (26.8%) are returning to the hospital within 30 days.

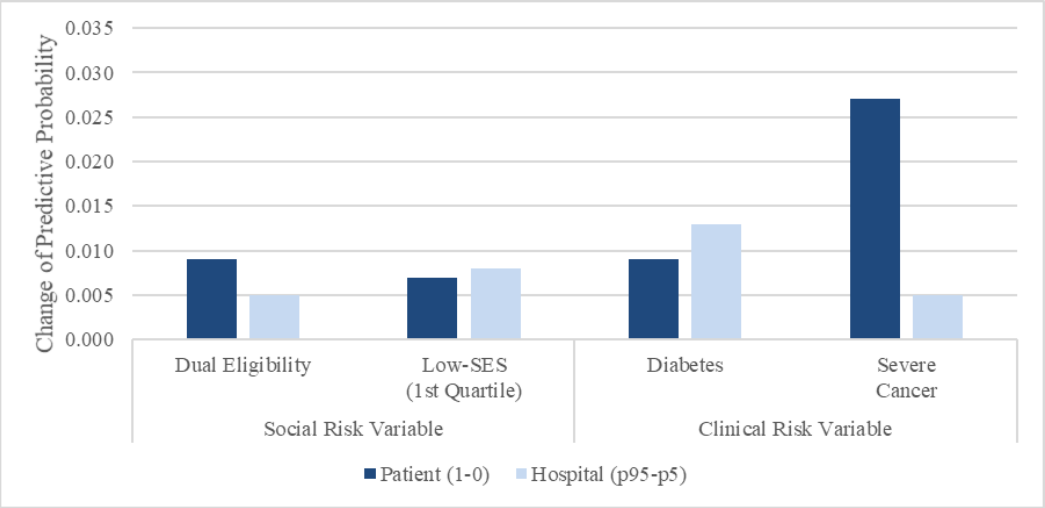
      Distribution of Hospital COPD RSRRs
      Dates of data: 07/2016-06/2019
      Number of Hospitals: 4,643
      Number of Admissions: 825,497
      Mean (SD): 19.6 (1)
      Range (Min-Max): 15.5-26.8
      Minimum: 15.5
      10^{th} percentile:18.5
      20^{th} percentile: 18.9
      30^{th} percentile: 19.2
      40^{th} percentile: 19.4
      50^{th} percentile: 19.6
      60^{th} percentile: 19.7
      70^{th} percentile: 19.9
      80^{th} percentile: 20.3
      90^{th} percentile: 20.8
      Maximum: 26.8
- **Issue 2: Use of VA Data.** One Panel member asked about the VA population, and if the developer had any comparisons or testing results that could demonstrate if the coefficients or models might be different in this population.
    - **Developer Response 2:** When VHA data were first added to the CMS readmission measures, extensive testing surrounding differences in cohorts, comorbidities and model fit were performed to determine if the addition of this unique population impacted the overall measure results. These analyses were shared with the VA and CMS, who determined that these measures could proceed with the VA population included for public reporting.
- **Issue 3: C-statistic.** One panel member noted that the c-statistic of 0.639 "may not be considered very good fit."

  - **Developer Response 3:** We ask that the Panel 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 comorbidities have a relatively limited relationship to the outcome 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. As noted by another SMP reviewer: "Model discrimination [~0.65] is acceptable for a readmission model; readmission models tend to have lower C stats compared to mortality models because more of the mechanism for readmission lie outside the hospital." We also note that the risk-decile plots show that the model performance is acceptable.
- **Issue 4: Approach to social risk factor adjustment.** One Panel member questioned if the developer was following NQF guidance in the approach to social risk factor adjustment, in that "if clinical risk factors explain all or most of the patient variation in the outcome, then NQF guidance does not support adding social risk factors that do not account for variation."
  - **Developer Response 4:** In NQF's 2017 publication, they cite in "Recommendation 5" that like clinical risk factors, social risk factors should "contribute unique variation in the outcome" and should not be redundant. We note that this measure was initially endorsed in 2008 and the model was developed initially without social risk factors and the clinical risk factors that are currently in the model were included based on criteria used at the time of development.

    Citations:

    National Quality Forum. Evaluation of the NQF Trial Period for risk adjustment for social risk factors, 2017. Accessed October 2, 2020; Available at: https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=85635

- **Issue 5: Patient- vs. hospital-level effects of social risk factors.** One Panel member noted that it appeared that the effects of each risk factor were the opposite of the developer's interpretation in the text.
  - **Developer Response 5:** Please note that because the AHRQ SES index is a continuous variable and the dual-eligibility variable is binary, you cannot directly compare the magnitudes of the coefficients in Table 8. Therefore, in order to quantitatively compare the relative size of the patient and hospital effects, we calculated a range of predicted probabilities of readmission based on the fitted model as described in 2b3.3a. The results shown in the testing attachment (Figure 3) and also shown below, are consistent with our written interpretation of the results regarding the patient and hospital-level effects of each variable.



- **Issue 6: Social risk factor adjustment interpretation of results**. One Panel member noted that the empiric data does not support the decision not to include adjustment for social risk factors.
    - **Developer Response 6**: The decision to risk adjust is based both on the empiric results (impact on model and measure scores), the conceptual model (hospitals are better able to mitigate the influence of social risk factors on the measured outcome than clinicians), and CMS' policy decision to adjust for dual eligibility at the program level (within the Hospital Readmission Reduction Program or HRRP). In HRRP, hospitals are stratified into peer groups by the proportion of dual eligible patients and are scored based on their performance within peer groups. Therefore, adjustment for social risk factors affects payment to providers, but not the public reporting of the quality measures. It is also consistent with Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation's (ASPE's) recommendation that quality measures that are used for public reporting should not be risk adjusted (ASPE 2020). As noted in our submission, CMS also confidentially reports disparities in the readmission measures to hospitals so that they have more detailed, actionable information about their patient population's social risk.

      Citation:

      Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed October 2, 2020.
- **Issue 7: Empiric validity.** A panel member wondered why the developer did not compare the COPD readmission measure to an excess days in acute care (EDAC) measure as they did for other readmission measures.
    - **Developer response 7:** There is no existing EDAC measure for COPD that the developer can use for comparison.

Measure #2515

Measure Title: Hospital 30-day all-cause risk-standardized readmission rate (RSRR) following Coronary artery bypass grafting (CABG) hospitalization

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Variation in the measure score:** A Panel member expressed concern about the low number of performance outliers. Another Panel member asked the developer to provide the distribution of measure scores.
    - **Developer Response 1:** Please note that performance categories are an implementation issue – CMS chooses to identify outliers based on 95% interval estimates, akin to 95% confidence intervals, which is a conservative approach to identifying performance outliers. These categories are also based on a comparison to the national average.
      The distribution of measure scores is shown below in deciles (this information was submitted in the portal, but perhaps was not visible to the reviewers). Note that the 10th and the 90th percentiles represent performance that is meaningfully different from average. Hospitals in the 10th percentile are performing about 13% better than average, and hospitals in the 90th percentile are performing about 13% worse than the average. In addition, the best-performing hospital has a readmission rate that is 62% lower than the worst-performing hospital. The median odds ratio is 1.23, meaning a patient has a 23% increase in the odds of a readmission at higher-risk hospital compared to a lower-risk hospital, therefore suggesting the measure can differentiate substantial variation in performance across hospitals.

Distribution of Hospital CABG RSRRs
Dates of data: 07/2016-06/2019
Number of Hospitals: 1,160
Number of Admissions: 13,1592
Mean(SD): 12.8(1.3)
Range(Min-Max): 8.6-22.6
Minimum: 8.6
10th percentile: 11.1
20th percentile: 11.7
30th percentile: 12.1
40th percentile: 12.5
50th percentile: 12.7
60th percentile: 13.0
70th percentile: 13.3
80th percentile: 13.8
90th percentile: 14.3
Maximum: 22.6

- **Issue 2: Increasing the minimum case requirement.** A Panel member noted that reliability might be improved if the minimum number of cases was substantially raised.
    - **Developer Response 2**: CMS currently uses a 25-case cut off for all the mortality and readmission measures. While we were not able to run additional reliability analyses with different case minimums within the response time for this feedback, this is an option we

are interested in exploring further. Note that there is a tradeoff between raising the case count and reducing the number of eligible hospitals for public reporting.

- **Issue 3: Insufficient testing for all-payer cohort:** Reviewers noted that there are insufficient testing results to support the separate all-payer cohort.
    - **Developer Response 3**: CORE has decided to change the measure specifications to limit the measures to the over 65, Medicare FFS patient population.

Validity

- **Issue 1: VA hospitals.** One Panel member asked about the VA population, and if the developer had any comparisons or testing results that would demonstrate if the coefficients or models might be different in this population.
    - **Developer Response 1:** When VHA data were first added to the CMS readmission measures, extensive testing surrounding differences in cohorts, comorbidities and model fit were performed to determine if the addition of this unique population impacted the overall measure results. These analyses were shared with the VA and CMS, who determined that these measures could proceed with the VA population included for public reporting.
- **Issue 2: Model validation.** Method panel members asked for additional information about how the developer validated its models with updated data.
    - **Developer Response 2:** Below, we have provided more context related to our standard approach for model validation and have addressed the specific comments from reviewers of other measures submitted by CORE that were shared with us via email. We use the same approach for validating models for all the measures we maintain.

      Overall, please consider evaluating these measures in the context of re-endorsement; we are building on our original measure development work and provide additional information based on current data; the results we present in our testing attachment are about the original model and how it performs with current data. The strength of the evidence we submitted in the testing attachments (risk-decile plots and c-statistic) supports that the models remain valid for use with current data.

      CORE's measures undergo an annual measure reevaluation process, which ensures that the risk-standardized mortality models are continually assessed and remain valid, given possible changes in clinical practice and coding standards over time. Modifications made to measure cohorts, risk models, and outcomes are informed by review of the most recent literature related to measure conditions or outcomes, feedback from various stakeholders, and empirical analyses, including assessment of coding trends that reveal shifts in clinical practice or billing patterns. Input is solicited from a workgroup composed of up to 20 clinical and measure experts, inclusive of internal and external consultants and subcontractors.

      We provide a link to the 2020 measure re-evaluation report for this measure. The report describes what CORE did for 2020 public reporting. We:

        - Updated the ICD-10 code-based specifications used in the measures. Specifically, we:
            - Incorporated the code changes that occurred in the FY 2019 version of the ICD-10-CM/PCS (effective with October 1, 2018+ discharges) into the cohort definitions and risk models; and,

    - Applied a modified version of the FY 2019 V22 CMS-Hierarchical Condition Category (HCC) crosswalk that is maintained by RTI International to the risk models.
  - Monitored code frequencies to identify any warranted specification changes due to possible changes in coding practices and patterns.
  - Evaluated the stability of the risk-adjustment model over the three-year measurement period by examining the model variable frequencies, model coefficients, and the performance of the risk-adjustment model in each year (July 2016-June 2017, July 2017-June 2018, and July 2018-June 2019).
  - For each of the conditions, we assessed logistic regression model performance in terms of discriminant ability for each year of data and for the three-year combined period. We computed two summary statistics to assess model performance: the predictive ability and the area under the receiver operating characteristic (ROC) curve (c-statistic).

  In summary:
  - We did not change which risk variables are in the model, but we updated the model coefficients using the updated dataset.
  - We do compare the model coefficients over time; as described above, we calculate model coefficients for each year of the three-year period, and for the three years combined. See Table 4.2.2 in the measure re-evaluation report.
  - Please note that risk-variable re-selection is part of our current workplan. To do this, we first needed a full three-year set of ICD-10-coded data, which we just reached this year. We had convened a Technical Expert Panel (TEP) for this work but COVID-19 hit just as we were about to start. We delayed the project to allow the TEP (which includes clinicians) to attend to pressing COVID-related matters. In addition, due to COVID, CORE analytic staff have, and continue to have, restricted access to the physical building where these types of analyses must occur.
- **Issue 3: C-statistic:** One panel member noted that the c-statistic "hovers around the mid-sixties, which makes one wonder overall goodness of fit of the model in terms of the included clinical and demographic factors."
  - **Developer Response 3:** We ask that the Panel 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 the outcome 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. As noted by another SMP reviewer: "Model discrimination [~0.65] is acceptable for a readmission model; readmission models tend to have lower C stats compared to mortality models because more of the mechanism for readmission lie outside the hospital."
- **Issue 4: Approach to social risk factor adjustment:** One Panel member questioned if the developer was following NQF guidance in the approach to social risk factor adjustment, in that

"if clinical risk factors explain all or most of the patient variation in the outcome, then NQF guidance does not support adding social risk factors that do not account for variation."

- **Developer Response 4:** In NQF's 2017 publication, they cite in "Recommendation 5" that like clinical risk factors, social risk factors should "contribute unique variation in the outcome" and should not be redundant.

  Citation:

  National Quality Forum. Evaluation of the NQF Trial Period for risk adjustment for social risk factors, 2017. Accessed October 2, 2020; Available at: https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=85635

- **Issue 5: Social risk factor adjustment interpretation of results**: One Panel member noted that the empiric data does not support the decision not to include adjustment for social risk factors.
  - **Developer Response 5**: The decision to risk adjust is based both on the empiric results (impact on model and measure scores), the conceptual model (hospitals are better able to mitigate the influence of social risk factors on the measured outcome than clinicians), and CMS' policy decision to adjust for dual eligibility at the program level (within the Hospital Readmission Reduction Program or HRRP). In HRRP, hospitals are stratified into peer groups by the proportion of dual eligible patients and are scored based on their performance within peer groups. Therefore, adjustment for social risk factors affects payment to providers, but not the public reporting of the quality measures. It is also consistent with the Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation's (ASPE's) recommendation that quality measures that are used for public reporting should not be risk adjusted (ASPE 2020). As noted in our submission, CMS also confidentially reports disparities in the readmission measures to hospitals so that they have more detailed, actionable information about their patient population's social risk.

    Citation:

    Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed October 2, 2020.

- **Issue 6: Empiric validity and face validity:** Panel members were concerned that the correlation with the CABG volume measure lower than they expected. In addition, a Panel member noted the absence of the results from the validation study with STS registry data and face validity results.
  - **Developer response 6**:

    Relationship of the outcome with volume: It is well established that CABG outcomes (mortality) have a weak relationship with volume (Marcin et al., 2008; Peterson et al., 2004), therefore we expected, and demonstrated, a weak, but significant correlation between the CABG readmission measure and CABG volume.

Validation study with STS registry data: Using STS CABG registry data and in collaboration with STS, we performed a clinical data validation study of the administrative cohort definition, risk adjustment model and hospital performance assessment that are outlined in detail in the original methodology report. The risk-adjustment validation produced a substantial correlation of RSRRs between the two measures in a matched cohort of patients, with an intraclass correlation coefficient of 0.92. When hospitals were categorized as "Better", "Worse" or "No different" than the national rate, over 97% (807 of 829) of hospitals in the matched cohort were categorized identically by the two measures (the vast majority were considered "No different than the national rate" by either measure).

Face Validity: During measure development we established the face validity of the measure. To systematically assess face validity, we surveyed the Technical Expert Panel (TEP) and asked each member to rate the following statement using a six-point scale (1=Strongly Disagree, 2=Moderately Disagree, 3=Somewhat Disagree, 4=Somewhat Agree, 5= Moderately Agree, and 6=Strongly Agree): "The readmission rates obtained from the readmission measure as specified will provide an accurate reflection of quality." Fourteen TEP members responded to the survey question as follows: Moderately Disagreed (2), Somewhat Disagreed (2), Somewhat Agreed (4), Moderately Agreed (5), and Strongly Agreed (1). Hence, 71% of TEP members agreed (43% moderately or strongly agreed) that the measure will provide an accurate reflection of quality.

Citations:

Marcin JP, Li Z, Kravitz RL, Dai JJ, Rocke DM, Romano PS. The CABG surgery volume-outcome relationship: temporal trends and selection effects in California, 1998-2004. *Health Serv Res*. 2008;43(1 Pt 1):174-192.

Peterson ED, Coombs LP, DeLong ER, Haan CK, Ferguson TB. Procedural Volume as a Marker of Quality for CABG Surgery. *JAMA.* 2004;291(2):195–201.

Other General Comments

- **Developer Comment:** We note that there is no clinical registry data in this measure, so any references to clinical registry data do not belong in the MIF for this measure.

Measure #2888

Measure Title: ACO Risk-Standardized Acute Admission Rates for Patients with Multiple Chronic Conditions

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: High signal to noise reliability.** One reviewer noted that the signal to noise ratios were high (0.96), and that they had not seen scores this high in prior measures. The reviewer asked for confirmation of the accuracy of the result. Another reviewer noted that it was "somewhat

surprising that the measure can include ACOs with only ONE patient attributed" and that they would like a further explanation of how "signal" comprises total variation in this measure (no "noise")?

- **Developer Response 1:** Thank you for your question. We have verified the accuracy of this value (0.96). This is the median reliability across all ACOs included in the measure (range: 0.123-0.997). The value is likely higher than reviewers typically see because this is an ACO-level measure, so there are many more patients per measured entity than for other types of providers (such as hospitals). Note that for this measure 2,515,727 patients with MCCs were attributed to just 559 ACOs; and while the algorithm uses a single patient for attribution, the median number of patients attributed to an ACO was 2,944 and the maximum was 34,858.

Validity

- **Issue 1**: **Qualifying conditions and validity**: One reviewer asked why the cohort qualifying conditions include certain chronic conditions related to the nervous, endocrine, cardiovascular, respiratory, and digestive systems but exclude other chronic conditions related to the integumentary, skeletal, muscular, lymphatic, urinary, and reproductive systems.
  - **Developer response 1**: We targeted applied transparent criteria with input from stakeholders, building on the NQF MCC framework to define criteria for cohort-defining chronic conditions. In brief: (1) we considered a condition "chronic" if it was persistent and had an adverse effect on health status, function, or quality of life (and was not an asymptomatic risk factor); (2) necessitated complex health care management, decision-making, or coordination when co-occurring with other conditions; and (3) high-quality ambulatory ACO care could lower the risk of admission among patients with these chronic conditions. Moreover, we empirically tested combinations of chronic conditions with respect to their prevalence and associated risk of admission and published the approach in the medical literature (Drye et al., 2018). Patients with other chronic conditions, such as those related to the integumentary, skeletal, muscular, lymphatic, urinary, and reproductive systems are not excluded but these conditions are not used as cohort-defining conditions.
    Citation:

    Drye, E. E. , Altaf, F. K. , Lipska, K. J. , Spatz, E. S. , Montague, J. A. , Bao, H. , Parzynski, C. S. , Ross, J. S. , Bernheim, S. M. , Krumholz, H. M. & Lin, Z. (2018). Defining Multiple Chronic Conditions for Quality Measurement. Medical Care, 56(2), 193–201.

- **Issue 2: Exclusions**: One reviewer asked for clarification if "10 days after discharge" was an exclusion per definition of denominator and that it was not listed in the exclusions. Another reviewer said they did not understand the reason for the 10-day buffer period after a discharge, and also noted that hospice was a concern.
  - **Developer Response 2:** Thank you for the opportunity to clarify this issue. The 10-days after discharge or "10-day buffer period" is a numerator (or outcome) exclusion but it also affects the denominator (person-time at risk). Any outcome within the 10-day buffer period is removed from the numerator. Persons are considered at risk for hospital admission if they are alive, enrolled in FFS Medicare, and not in the hospital

during the measurement period. In addition to time spent in the hospital, we also exclude from at-risk time: 1) time spent in a SNF or acute rehabilitation facility; 2) the time within 10 days following discharge from a hospital, SNF, or acute rehabilitation facility; and 3) time after entering hospice care. Note that the patient is not removed from the denominator, we are just subtracting the 10-days of person-time.

The 10-day buffer period (10 days following discharge from a hospital) is a period of transition back to community-based care, and other factors in addition to ambulatory care, including care received in the hospital and post-discharge planning, contribute to the risk of admission; therefore, the measure does not hold clinicians accountable for admissions in this timeframe. This buffer period allows time for patients to be seen within 7 days of discharge as recommended in CMS's Transitional Care Management (TCM) service guidelines and for the ambulatory care provider's care plan to take effect. CMS's TCM service guidelines encourage providers to have a face-to-face visit within 7 days of discharge for Medicare patients with high medical decision complexity.
The measure also excludes admissions that occur after the patient has entered hospice. Once a patient enters hospice care, a goal of care is to prevent the need for hospital care. However, ambulatory care providers may be attributed the patient and have relatively little influence on end-of-life care once a patient is enrolled in hospice and served by a hospice team.

- **Issue 3: Dates of Data**: One reviewer asked if we used 2012 data for testing.
    - **Developer Response 3**: As described in Section 1.2 of the testing attachment, the cohort and outcome were derived from Medicare Fee-for-Service (FFS) administrative claims data (Parts A and B) and the Medicare Enrollment Database (EDB) from calendar year (CY) 2018. None of the data we used for testing was from 2012.
- **Issue 4: R-squared value:** One reviewer cited concern that the deviance of R-squared of 11% was not "good enough"
    - **Developer Response 4:** Thank you for your comment. Please note that another SMP reviewer pointed out that while this is low, it is consistent with other risk adjustment models that have been endorsed by NQF.
- **Issue 5**: **Social risk factor adjustment**: One reviewer noted that CORE's "own results show that SES should be a risk factor in the model" and another reviewer noted that the rationale for risk adjusting this ACO MCC measure was in conflict with the rationale presented by CORE for hospital-level readmission measures going through NQF endorsement.
    - **Developer Response 5:** We thank the reviewer for their comments. Please note that CMS did decide to include two social risk factors (low physician-specialist density, and low AHRQ SES index) in both the MIPS and ACO MCC measures, however the inclusion of these two social risk factors in the ACO MCC updated measure was done primarily to fully align the measure with the MIPS clinician-group level MCC measure. As discussed in the MIPS MCC and ACO measure NQF applications, CMS took a wholistic look, both quantitative and qualitative, at the benefits and risks of adjusting for social risk factors in each of the measures and balances sometimes competing considerations in reaching a decision on whether to include social risk factor adjustment in the models. CMS included the AHRQ SES index and physician-specialist density variables in the MIPS MCC measure because MIPS-eligible clinicians working in the community may have a limited

ability to influence these community-based contextual factors that affect admission risk. However, CMS acknowledges that ACOs typically have a greater ability to mitigate the increased admission risk associated with social risk factors than smaller clinician groups that will be measured by the MIPS MCC measure. On balance, however, CMS decided it was programmatically very important to fully align the two admission measures' specifications given their use in the two QPP programs and the possibility that individual providers and groups may be assessed by either or both measures over time given their financial arrangements with clinician groups and ACOs.

The reviewer is right to point out that while CMS did decide to add low AHRQ SES index to the risk adjustment model for the clinician-group-level MCC measure, they did not adjust the hospital-level readmission measures. CMS decisions were informed by the conceptual models for each measure that specify the pathways through which social risk factors influence the measured outcome. CMS also considered the literature and results of empiric analyses.

The difference in the approach to the hospital readmission and MIPS/ACO admission measures is driven largely by the different context in which the outcomes are being measured. The ACO and MIPS MCC measures assess clinicians' ability to manage patients over the course of a year in the ambulatory setting in a way that minimizes acute unplanned admissions. The conceptual model for these measures acknowledges that social risk factors influence the outcome, and clinicians in the community may not be able to fully mitigate the pathways through which this influence occurs throughout the course of a year given their limited interactions with their patients. In contrast, the hospital readmission measures assess a near-term (30-day) outcome of an acute episode of care, during which generally well-resourced institutions have the patient under their care and can identify, plan for and mitigate specific factors that put patients at increased risk of readmission. In other words, hospitals have a greater ability to mitigate the influence of social risk factors on the measured outcome. In fact, quality improvement efforts to reduce 30-day hospital readmissions focus on those very areas that hospitals can control, such as improving care transitions, including patient education, medication reconciliation, discharge planning, and coordination of outpatient care, which have been shown to reduce readmission rates.

Other General Comments

- **Issue 1: Number of ACOs:** One reviewer asked the developer to clarify the number of ACOs tested. The reviewer stated that the number in the MIF was 114 but the number in the testing attachment is 559.

    - **Developer Response 1:** The correct number is 559.

Subgroup 2

Measure #0229

Measure Title: Hospital 30-day, All-Cause, Risk-Standardized Mortality Rate (RSMR) following heart failure (HF) hospitalization

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Insufficient testing for all-payer cohort:** Reviewers noted that there are insufficient testing results to support the separate all-payer cohort.
    - **Developer Response 1**: CORE has decided to change the measure specifications to limit the measure to the over 65, Medicare FFS patient population.
- **Issue 2: Measure specifications**. One Panel member asked if the measure includes both patients who were discharged alive and patients who were discharged dead?
    - **Developer Response 2**: Yes, in-hospital deaths are part of the outcome.
- **Issue 3: 25-case threshold.** Please explain how the 25-case minimum was established
    - **Developer Response 3:** The 25-case minimum was established when these outcome measures were first implemented in public reporting. CMS chose a cutoff of 25 eligible cases to align with the minimum volume requirement used in the publicly reported process of care measures.

Validity

- **Issue 1: Model validation.** Method panel members asked for additional information about how the developer validated its models with updated data.
    - **Developer Response 1:** Below we have provided more context related to our standard approach for model validation, and have addressed the specific comments from reviewers of other measures submitted by CORE that were shared with us via email. We use the same approach for validating models for all of the measures we maintain.

      Overall, please consider evaluating these measures in the context of re-endorsement; we are building on our original measure development work and provide additional information based on current data; the results we present in our testing attachment are about the original model and how it performs with current data. The strength of the evidence we submitted in the testing attachments (risk-decile plots and c-statistic) supports that the models remain valid for use with current data.

      CORE's measures undergo an annual measure reevaluation process, which ensures that the risk-standardized mortality models are continually assessed and remain valid, given possible changes in clinical practice and coding standards over time. Modifications made to measure cohorts, risk models, and outcomes are informed by review of the most recent literature related to measure conditions or outcomes, feedback from various stakeholders, and empirical analyses, including assessment of coding trends that reveal shifts in clinical practice or billing patterns. Input is solicited from a workgroup composed of up to 20 clinical and measure experts, inclusive of internal and external consultants and subcontractors.

      We provide a link to the 2020 measure re-evaluation report for this measure. The report describes what CORE did for 2020 public reporting; we:

        - Updated the ICD-10 code-based specifications used in the measures. Specifically, we:

- Incorporated the code changes that occurred in the FY 2019 version of the ICD-10-CM/PCS (effective with October 1, 2018+ discharges) into the cohort definitions and risk models; and,
- Applied a modified version of the FY 2019 V22 CMS-Hierarchical Condition Category (HCC) crosswalk that is maintained by RTI International to the risk models.
- Monitored code frequencies to identify any warranted specification changes due to possible changes in coding practices and patterns;
- Evaluated the stability of the risk-adjustment model over the three-year measurement period by examining the model variable frequencies, model coefficients, and the performance of the risk-adjustment model in each year (July 2016-June 2017, July 2017-June 2018, and July 2018-June 2019).
- For each of the conditions, we assessed logistic regression model performance in terms of discriminant ability for each year of data and for the three-year combined period. We computed two summary statistics to assess model performance: the predictive ability and the area under the receiver operating characteristic (ROC) curve (c-statistic).

In summary:

- We did not change which risk variables are in the model, but we updated the model coefficients using the updated dataset.
- We do compare the model coefficients over time; as described above, we calculate model coefficients for each year of the three-year period, and for the three years combined. See Table 4.4.2 in the measure re-evaluation report.
- Please note that risk-variable re-selection is part of our current workplan. To do this, we first needed a full three-year set of ICD-10-coded data, which we just reached this year. We had convened a Technical Expert Panel (TEP) for this work but COVID hit just as we were about to start. We delayed the project to allow the TEP (which includes clinicians) to attend to pressing COVID-related matters. In addition, due to COVID, CORE analytic staff have, and continue to have, restricted access to the physical building where these types of analyses must occur.

- **Issue 2: Model odds ratios.** Interpretation of odd ratios; especially for factors associated with lower risk of mortality. Could some of these 'protective' factors be due to collinearity with other risk-factors? Were results assessed for clinical plausibility?
    - **Developer Response 2**: The selection of risk factors for these measures is based on empirical analysis, prior literature, and clinical judgment. The process allows for the selection of variables that are protective in relation to the outcome if they were shown to be significant or of clinical importance. We do test for collinearity with variables that may be clinically similar, and in those instances we often combine similar conditions into one risk variable. Risk models are also reviewed for clinical sensibility, however administrative claims data do not always function identically to clinical variables.
- **Issue 3: Social risk factor analyses: net reclassification:** Two reviewers asked the developer if they had examined the reclassification of hospitals following the addition of risk variables to the model.

- **Developer Response 3:** We did not perform a reclassification analysis. Note that patients with either social risk factor have a lower observed mortality rate, and odds ratios for patients with social risk factors are less than one, therefore any adjustment would be in the opposite direction than what has been the expressed concern of stakeholders interested in adding such adjustment to the models. We agree, however, that model performance should not be judged based on the c-statistic alone. In the context of evaluating hospital performance, it is relevant to assess the impact on hospital performance. We have looked at both the net reclassification index (NRI) and the integrated discrimination improvement (IDI) a few years ago in our development work. We have decided not to use them given some important findings and concerns raised by others.

  Citations:

  Pepe M, Fan J, Feng Z, Gerds T, Hilden H (2015) The Net Reclassification Index (NRI): A misleading measure of prediction improvement even with independent test data sets. Stat Biosci 7:282-295

  Hilder J, Gerds T (2013) A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med 33(19):3405-14.

- **Issue 4: Empiric validity, comparator measures**: Panel members expressed concern about the measures chosen as comparator measures for empiric validation.
  - **Developer Response 4**: CORE met with measure experts and clinicians to identify candidate comparator measures for validation. Unfortunately there are few available relevant measures and even fewer relevant measures with publicly available data for us to use for such analyses, particularly of the process type that would be related to the measure's focus.

Other General Comments

- **Issue 1:** One Panel member noted that we referred to a "decomposition" analysis in our assessment of the impact of adding social risk factors, but that the reviewer was unable to find the results.
  - **Developer Response 1:** Due to an editing error, the reference to the decomposition analysis was in our submission but we did not run this analysis for this measure.

Measure #0230

Measure Title: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization for patients 18 and older

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Split sample and signal-to-noise reliability**. A few Panel members disagreed with the developer's interpretation of the reliability testing results.

- **Developer Response 1:** The split-sample reliability using two randomly split, non-overlapping samples, was 0.428. (Note the signal-to-noise reliability for this measure was 0.59; split-sample reliability tends to provide a lower estimate than signal-to-noise, in general [Yu et al., 2013].) We interpret the split-sample reliability result as "moderate" based on the standards established by Landis and Koch (1977). This result is consistent with those from known, familiar, and related contexts. The current context is measuring provider quality, or specifically provider propensity to provide appropriate care as measured by subsequent outcomes. We identified several studies, which we think support the Landis & Koch guidelines when assessing test-retest reliability in the context of hospital measurement.
- Hall et al calculated test-retest reliability for determining comorbidities from chart abstraction (Hall et al., 2006). In this study, multiple abstracters abstracted the same charts and the results were used to calculate four different common comorbidity scores. For three of the indices, test-retest reliabilities ranged from 0.59-0.68, with the fourth (the Charlson comorbidity score) achieving 0.80. We would argue that chart abstraction, with test-retest reliabilities in the 'moderate' to 'substantial' range, should be inherently more reliable than measuring hospital quality.
- Cruz et al report reliabilities for collecting risk factor information from patients presenting to an emergency department with potential acute coronary syndrome (ACS) (Cruz et al., 2009). Each patient was queried twice, once by a clinician and once by a trained research assistant, and the reliabilities for a range of risk factors were calculated; these ranged from 0.28 (associated symptoms) to 0.69 (cardiac risk factors), with all other factors in the 0.30-0.56 range.
- Hand et al report test-retest reliabilities for bedside clinical assessment of suspected stroke (Hand et al., 2006). Pairs of observers independently assessed suspected stroke patients; findings were recorded on a standard form to promote consistency. The reliabilities were calculated for the full range of diagnostic factors: for vascular factors reliabilities ranged from 0.47-0.69 with only four of eight above 0.6; for history they ranged from 0.37-0.65 with only five of 12 above 0.6; other categories were similar (though reliability=1 for whether the patients were conscious).

  Citations:

  Hall SF, Groome PA, Streiner DL, Rochon PA. Interrater reliability of measurements of comorbid illness should be reported. J Clin Epidemiol. 2006 Sep;59(9):926-33.

  Cruz CO, Meshberg EB, Shofer FS, McCusker CM, Chang AM, Hollander JE. Interrater reliability and accuracy of clinicians and trained research assistants performing prospective data collection in emergency department patients with potential acute coronary syndrome. Ann Emerg Med. 2009 Jul;54(1):1-7.

  Hand PJ, Haisma JA, Kwan J, Lindley RI, Lamont B, Dennis MS, Wardlaw JM. Interobserver agreement for the bedside clinical assessment of suspected stroke. Stroke. 2006 Mar;37(3):776-80.

Yu H, Mehrotra A, Adams J. Reliability of utilization measures for primary care physician profiling. Healthcare,2013, 1:22-29.

- **Issue 2: 25-case threshold.** Please explain how the 25-case cut off was established
    - **Developer Response 2:** The 25-case minimum was established when these outcome measures were first implemented in public reporting. CMS chose a cutoff of 25 eligible cases to align with the minimum volume requirement used in the publicly reported process of care measures.
- **Issue 3: Increasing the minimum case requirement.** A Panel member noted that reliability might be improved if the minimum number of cases was substantially raised.
    - **Developer Response 3**: CMS currently uses a 25-case cut-off for all mortality and readmission measures. While we were not able to run additional reliability analyses with different case minimums within the response time for this feedback, this is an option we are interested in exploring further. Note that there is a tradeoff between raising the case count and reducing the number of eligible hospitals for public reporting.
- **Issue 4: Insufficient testing for all-payer cohort:** Reviewers noted that there are insufficient testing results to support the separate all-payer cohort.
    - **Developer Response 4**: CORE has decided to change the measure specifications to limit the measures to the over 65, Medicare FFS patient population.
- **Issue 5: Measure specifications**. One Panel member asked if the measure includes both patients who were discharged alive and patients who were discharged dead?
    - **Developer Response 5**: Yes, in-hospital deaths are part of the outcome.

Validity

- **Issue 1: Low numbers of outliers identified by performance categories:** A Panel member expressed concern about the relatively low proportion of performance outliers.
    - **Developer Response 1:** Please note that performance categories are an implementation issue – CMS chooses to identify outliers based on 95% interval estimates, akin to 95% confidence intervals, which is a conservative approach to identifying performance outliers. We note that the median odds ratio suggests a meaningful increase in the risk of mortality if a patient is admitted with AMI at a higher risk hospital compared to a lower risk hospital. A value of 1.19 indicates that a patient has a 19% increase in the odds of mortality at higher-risk hospital compared to a lower-risk hospital, indicating that the measure can identify meaningful differences in hospital performance.
- **Issue 2: Validating the claims-based model**. Panel members commented on the validity of the measure.
    - **Developer Response 2:** CORE is providing additional information that was not included in the submission. During measure development CORE validated the performance of the administrative and medical record models is similar. The areas under the receiver operating characteristic (ROC) curve are 0.69 and 0.77, respectively, for the two models. We estimated hospital-level RSMRs using the corresponding hierarchical logistic regression administrative and medical record models for the linked patient sample. We then examined the linear relationship between the two sets of estimates using regression techniques and weighting by the total number of cases in each hospital. The

correlation coefficient of the standardized rates from the administrative and medical record models is 0.91 which shows that there was a strong correlation in rates calculated from the clinical and admin models.

Citation:

Krumholz HM, Wang Y, Mattera JA, Wang Y, Han LF, Ingber MJ, Roman S, Normand SL. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation. 2006 Apr 4;113(13):1683-92.

- **Issue 3: Model validation.** Method panel members asked for additional information about how the developer validated its models with updated data.
    - **Developer Response 3:** Below we have provided more context related to our standard approach for model validation, and have addressed the specific comments from reviewers of other measures submitted by CORE that were shared with us via email. We use the same approach for validating models for all of the measures we maintain.

      Overall, please consider evaluating these measures in the context of re-endorsement; we are building on our original measure development work and provide additional information based on current data; the results we present in our testing attachment are about the original model and how it performs with current data. The strength of the evidence we submitted in the testing attachments (risk-decile plots and c-statistic) supports that the models remain valid for use with current data.

      CORE's measures undergo an annual measure reevaluation process, which ensures that the risk-standardized mortality models are continually assessed and remain valid, given possible changes in clinical practice and coding standards over time. Modifications made to measure cohorts, risk models, and outcomes are informed by review of the most recent literature related to measure conditions or outcomes, feedback from various stakeholders, and empirical analyses, including assessment of coding trends that reveal shifts in clinical practice or billing patterns. Input is solicited from a workgroup composed of up to 20 clinical and measure experts, inclusive of internal and external consultants and subcontractors.

      We provide a link to the 2020 measure re-evaluation report for this measure. The report describes what CORE did for 2020 public reporting; we:

        - Updated the ICD-10 code-based specifications used in the measures. Specifically, we:
        - Incorporated the code changes that occurred in the FY 2019 version of the ICD-10-CM/PCS (effective with October 1, 2018+ discharges) into the cohort definitions and risk models; and,
        - Applied a modified version of the FY 2019 V22 CMS-Hierarchical Condition Category (HCC) crosswalk that is maintained by RTI International to the risk models.
        - Monitored code frequencies to identify any warranted specification changes due to possible changes in coding practices and patterns;
        - Evaluated the stability of the risk-adjustment model over the three-year measurement period by examining the model variable frequencies, model

coefficients, and the performance of the risk-adjustment model in each year (July 2016-June 2017, July 2017-June 2018, and July 2018-June 2019).

- For each of the conditions, we assessed logistic regression model performance in terms of discriminant ability for each year of data and for the three-year combined period. We computed two summary statistics to assess model performance: the predictive ability and the area under the receiver operating characteristic (ROC) curve (c-statistic).

In summary:

- We did not change which risk variables are in the model, but we updated the model coefficients using the updated dataset.
- We do compare the model coefficients over time; as described above, we calculate model coefficients for each year of the three-year period, and for the three years combined. See Table 4.2.2 in the measure re-evaluation report.
- Please note that risk-variable re-selection is part of our current workplan. To do this, we first needed a full three-year set of ICD-10-coded data, which we just reached this year. We had convened a Technical Expert Panel (TEP) for this work but COVID hit just as we were about to start. We delayed the project to allow the TEP (which includes clinicians) to attend to pressing COVID-related matters. In addition, due to COVID, CORE analytic staff have, and continue to have, restricted access to the physical building where these types of analyses must occur.

- **Issue 4: Model odds ratios.** Interpretation of odd ratios; especially for factors associated with lower risk of mortality. Could some of these 'protective' factors be due to collinearity with other risk-factors? Were results assessed for clinical plausibility? The reviewer also asked if the acute myocardial infarction (AMI) variable was supposed to be a constant, or if it describes a history of AMI.
    - **Developer Response 4**: The selection of risk factors for these measures is based on empirical analysis, prior literature, and clinical judgment. The process allows for the selection of variables that are protective in relation to the outcome if they were shown to be significant or of clinical importance. We do test for collinearity with variables that may be clinically similar, and in those instances we often combine similar conditions into one risk variable. Risk models are also reviewed for clinical sensibility, however administrative claims data do not always function identically to clinical variables. The risk models have been validated against medical record abstracted risk models data and perform similarly. Please note that the AMI variable does describe a history of AMI.
- **Issue 5: Social risk factor adjustment.** One Panel member noted an editing error in our submission that stated "the relationship between dual-eligible status and AHRQ low SES is in the opposite direction [lower odds ratios] than what has been the expressed concern of stakeholders interested in adding such adjustment to the models."
    - **Developer Response 5:** The reviewer is correct, the odds ratios are higher, not lower. Due to an editing error this sentence was accidently incorporated into the text.
- **Issue 6: Minimum case threshold**. One Panel member asked if the developers assess the minimum threshold of cases per hospital needed to achieve a mean reliability of 0.7?

  - **Developer Response 6**: No, we did not. This measure has been in public reporting since 2008, with the same 25 case threshold. The 25-case minimum was established when these outcome measures were first implemented in public reporting. CMS chose a cutoff of 25 eligible cases to align with the minimum volume requirement used in the publicly reported process of care measures.
- **Issue 7: Empiric validity, comparator measures**: Panel members expressed concern about the measures chosen as comparator measures for empiric validation. Some panel members wished the developer has used process measures that would be expected to be related to the outcome. One Panel member noted that there was no attempt to correlate with other valid measures, such as the American College of Cardiology hospital MI mortality metric.
  - **Developer Response 7**: CORE met with measure experts and clinicians to identify candidate comparator measures for validation. Unfortunately, there are few available relevant measures and even fewer relevant measures with publicly available data for us to use for such analyses, particularly of the process type that would be related to the measure's focus. We did assess the feasibility of using other mortality measures that were reported for all or most acute care hospitals and failed to find others that are readily available at the hospital level for analysis.
- **Issue 8**: **Empiric validity, p-values and confidence intervals**. One panel member requested p-values and confidence intervals for the correlation analyses that were performed.
  - **Developer Response 8**: P-values and confidence intervals for each correlation analysis are shown in the table below.

Analysis	Correlation with RSMR	95% Lower Limit	95% Upper Limit	p-value
Correlation between RSMR and Star Rating Standardized Mortality Ratings	-0.409	-0.435	-0.382	<.0001
Correlation between RSMR and Star Rating Standardized Summary Ratings	-0.204	-0.233	-0.174	<.0001

Other General Comments

- **Issue 1:** One Panel member noted that we referred to a "decomposition" analysis in our assessment of the impact of adding social risk factors, but that the reviewer was unable to find the results.
  - **Developer Response 1:** Due to an editing error, the reference to the decomposition analysis was in our submission but we did not run this analysis for this measure.

Measure #0468

Measure Title: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following pneumonia hospitalization

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Insufficient testing for all-payer cohort:** One Panel member noted that there are insufficient testing results to support the non-FFS cohort.
    - **Developer Response 1**: CORE has decided to change the measure specifications to limit the measure to the over 65, Medicare FFS patient population.
- **Issue 2: Unclear specification of the outcome**. One Panel member asked if the outcome includes both patients who were discharged alive and patients who were discharged dead.
    - **Developer Response 2**: Yes, in-hospital deaths are part of the outcome. We will edit the description as the reviewer suggested.

Validity

- **Issue 1: 25-case threshold.** Please explain how the 25-case cut off was established
    - **Developer Response 1:** The 25-case minimum was established when these outcome measures were first implemented in public reporting. CMS chose a cutoff of 25 eligible cases to align with the minimum volume requirement used in the publicly reported process of care measures.
- **Issue 2: Model validation.** Method panel members asked for additional information about how the developer validated its models with updated data.
    - **Developer Response 2:** Below we have provided more context related to our standard approach for model validation, and have addressed the specific comments from reviewers of other measures submitted by CORE that were shared with us via email. We use the same approach for validating models for all of the measures we maintain.

      Overall, please consider evaluating these measures in the context of re-endorsement; we are building on our original measure development work and provide additional information based on current data; the results we present in our testing attachment are about the original model and how it performs with current data. The strength of the evidence we submitted in the testing attachments (risk-decile plots and c-statistic) supports that the models remain valid for use with current data.

      CORE's measures undergo an annual measure reevaluation process, which ensures that the risk-standardized mortality models are continually assessed and remain valid, given possible changes in clinical practice and coding standards over time. Modifications made to measure cohorts, risk models, and outcomes are informed by review of the most recent literature related to measure conditions or outcomes, feedback from various stakeholders, and empirical analyses, including assessment of coding trends that reveal shifts in clinical practice or billing patterns. Input is solicited from a workgroup composed of up to 20 clinical and measure experts, inclusive of internal and external consultants and subcontractors.

      We provide a link to the 2020 measure re-evaluation report for this measure. The report describes what CORE did for 2020 public reporting; we:
        - Updated the ICD-10 code-based specifications used in the measures. Specifically, we:

- Incorporated the code changes that occurred in the FY 2019 version of the ICD-10-CM/PCS (effective with October 1, 2018+ discharges) into the cohort definitions and risk models; and,
- Applied a modified version of the FY 2019 V22 CMS-Hierarchical Condition Category (HCC) crosswalk that is maintained by RTI International to the risk models.
- Monitored code frequencies to identify any warranted specification changes due to possible changes in coding practices and patterns;
- Evaluated the stability of the risk-adjustment model over the three-year measurement period by examining the model variable frequencies, model coefficients, and the performance of the risk-adjustment model in each year (July 2016-June 2017, July 2017-June 2018, and July 2018-June 2019).
- For each of the conditions, we assessed logistic regression model performance in terms of discriminant ability for each year of data and for the three-year combined period. We computed two summary statistics to assess model performance: the predictive ability and the area under the receiver operating characteristic (ROC) curve (c-statistic).

In summary:

- We did not change which risk variables are in the model, but we updated the model coefficients using the updated dataset.
- We do compare the model coefficients over time; as described above, we calculate model coefficients for each year of the three-year period, and for the three years combined. See Table 4.5.2 in the measure re-evaluation report.
- Please note that risk-variable re-selection is part of our current workplan. To do this, we first needed a full three-year set of ICD-10-coded data, which we just reached this year. We had convened a Technical Expert Panel (TEP) for this work but COVID hit just as we were about to start. We delayed the project to allow the TEP (which includes clinicians) to attend to pressing COVID-related matters. In addition, due to COVID, CORE analytic staff have, and continue to have, restricted access to the physical building where these types of analyses must occur.

- **Issue 3: Model odds ratios.** One Panel member asked about our interpretation of odd ratios; especially for factors associated with lower risk of mortality. The reviewer asked: "Could some of these 'protective' factors be due to collinearity with other risk-factors? Were results assessed for clinical plausibility?"
    - **Developer Response 3**: The selection of risk factors for these measures is based on empirical analysis, prior literature, and clinical judgment. The process allows for the selection of variables that are protective in relation to the outcome if they were shown to be significant or of clinical importance. Risk models are also reviewed for clinical sensibility, however administrative claims data do not always function identically to clinical variables. The risk models have been validated against medical record abstracted risk model data and perform similarly. We do test for collinearity with variables that may be clinically similar, and in those instances we often combine similar conditions into one risk variable.

- **Issue 4: Social risk factor adjustment, net reclassification index.** One Panel member commented that the developers should consider an analysis such as net reclassification index or some other measure to guide the decision of adjusting the measure for social risk.
    - **Developer Response 4**: We agree that model performance should not be judged based on the c-statistic alone. In the context of evaluating hospital performance, it is relevant to assess the impact on hospital performance. We have looked at both the net reclassification index (NRI) and the integrated discrimination improvement (IDI) a few years ago in our development work. We have decided not to use them given some important findings and concerns raised by others.

      Citations:

      Pepe M, Fan J, Feng Z, Gerds T, Hilden H (2015) The Net Reclassification Index (NRI): A misleading measure of prediction improvement even with independent test data sets. Stat Biosci 7:282-295
      Hilder J, Gerds T (2013) A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med 33(19):3405-14.
- **Issue 5: Social risk factor adjustment, clarification.** One Panel member noted an editing error in our submission that stated "the relationship between dual-eligible status and AHRQ low SES is in the opposite direction [lower odds ratios] than what has been the expressed concern of stakeholders interested in adding such adjustment to the models."
    - **Developer Response 5:** The reviewer is correct, the odds ratios are higher, not lower. Due to an editing error this sentence was accidently incorporated into the text. The sentence will be removed.
- **Issue 6: Empiric validity.** Panel members expressed concern about the measures chosen as comparator measures for empiric validation. Some panel members wished the developer has used process measures that would be expected to be related to the outcome.
    - **Developer Response 6:** CORE met with measure experts to identify candidate comparator measures for validation. Unfortunately there are few available relevant measures and even fewer relevant measures with publicly available data for us to use for such analyses, particularly of the process type that would be related to the measure's focus.

Other General Comments

- **Issue 1:** One Panel member noted that we referred to a "decomposition" analysis in our assessment of the impact of adding social risk factors, but that the reviewer was unable to find the results.

    - **Developer Response 1:** Due to an editing error, the reference to the decomposition analysis was in our submission but we did not run this analysis for this measure.

Measure #1550

Measure Title: Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Insufficient testing for all-payer cohort**: Panel members noted that there are insufficient testing results to support the all-payer cohorts.
    - **Developer Response 1:** CORE has decided to change the measure specifications to limit the measure to the over 65, Medicare FFS patient population.

- **Issue 2: Measure specifications, 25-case minimum.** Please explain how the 25-case cut off was established
    - **Developer Response 2:** The 25-case minimum was established when these outcome measures were first implemented in public reporting. CMS chose a cutoff of 25 eligible cases to align with the minimum volume requirement used in the publicly reported process of care measures.

Validity

- **Issue 1: Model odds ratios.** One panel member asked about the Interpretation of odd ratios, especially for factors associated with lower risk of readmission. Could some of these 'protective' factors be due to collinearity with other risk-factors? Were results assessed for clinical plausibility?
    - **Developer Response 1**: The selection of risk factors for these measures is based on empirical analysis, prior literature, and clinical judgment. The process allows for the selection of variables that are protective in relation to the outcome if they were shown to be significant or of clinical importance. We do test for collinearity with variables that may be clinically similar, and in those instances we often combine similar conditions into one risk variable. Risk models are also reviewed for clinical sensibility, however administrative claims data do not always function identically to clinical variables. The risk model have been validated against medical record abstracted risk models data and perform similarly.
- **Issue 2: Social risk factor adjustment interpretation of results**: One Panel member noted that the empiric data does not support the developer's decision not to include adjustment for social risk factors.
    - **Developer Response 2**: The decision to risk adjust is based both on the empiric results (impact on model and measure scores) and the conceptual model (hospitals are better able to mitigate the influence of social risk factors on the measured outcome than clinicians). It is also consistent with Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation's (ASPE's) recommendation that quality measures that are used for public reporting should not be risk adjusted (ASPE 2020).

      Citation:

      Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed October 2, 2020.

- **Issue 3: C-statistic**: One panel member noted that the discrimination of the model is weak, suggesting that there are factors relevant to the outcome of interest that are not accounted for in the model.
    - **Developer Response 3**: We ask that the Panel 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 comorbidities have a relatively limited relationship to the occurrence the outcome 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.
- **Issue 4: Empiric validity.** Panel members expressed concern about the measures chosen as comparator measures for empiric validation. Some panel members wished the developer has used process measures that would be expected to be related to the outcome, or outcome measures related to functional status. A panel member suggested we look at the correlation with mortality.
    - **Developer Response 4:** CORE met with measure experts to identify candidate comparator measures for validation. Unfortunately there are few available relevant measures and even fewer relevant measures with publicly available data for us to use for such analyses, particularly of the type that would be related to the measure's focus. Note that mortality is counted as a complication for this measure (death during the index admission or within 30 days of the index admission).

      While complications as a quality measure has its own inherent face validity, we also performed a validation study against medical records, described in the original technical report. The study included 644 patients - 319 patients who the claims-based measure identified as having one or more complications and 325 who the measure identified as having no complications. The medical record acquisition rate for these 644 patients was 96% (644 patient records received / 674 patient records requested). Overall measure agreement was 93% (598/644 patients). We made changes to the measure based on the validation study (described in detail in the original technical report); after the proposed measure changes were implemented, measure agreement between claims data and the medical record will increase to 99% (635/644 patients).
- **Issue 5: Inability to distinguish between sites with multiple complication**. One panel member expressed concern that patients with more than one complication were not differentiated from those with one complication, since multiple complications are associated with higher mortality.
    - **Developer response 5**: We wanted to develop the measure using a dichotomous yes/no for the outcome, therefore we do not count the second complication as an event. This was a trade-off between capturing a relevant outcome and complicating the construction of the measure.
- **Issue 6: Variation in the measure score:** A Panel member expressed concern about the relatively low proportion of performance outliers. Another panel member expressed concern about the tight distribution of measure scores.
    - **Developer Response 6:** Please note that performance categories are an implementation issue – CMS chooses to identify outliers based on 95% interval estimates, akin to 95%

confidence intervals, which is a conservative approach to identifying performance outliers.

The distribution of measure scores in deciles is shown below. There are meaningful differences in the distribution – for example, hospitals in the 10th percentile are performing about 24% better than the average performer, and hospitals in the 90th percentile are performing about 20% worse than the average performer.

In addition, the median odds ratio (1.38) suggests a meaningful increase in the risk of complications if a patient has a THA/TKA procedure at a higher-risk hospital compared to a lower-risk hospital. A value of 1.38 indicates that a patient has a 38% increase in the odds of a complications at a higher-risk hospital compared to a lower-risk hospital, indicating the impact of quality on the outcome rate. This variation suggests there remain differences in the quality of care received across hospitals for THA/TKA procedures. This evidence supports continued measurement to reduce the variation.

Distribution of Hospital THA/TKA RSCRs over Different Time Periods
Number of Hospitals: 3418
Number of Admissions: 962,744
Mean (SD): 2.5(0.5)
Range (Min-Max): 1.2-10.6
Minimum: 1.2
10th percentile: 1.9
20th percentile: 2.1
30th percentile: 2.3
40th percentile: 2.3
50th percentile: 2.4
60th percentile: 2.5
70th percentile: 2.6
80th percentile: 2.8
90th percentile: 3.0
Maximum: 10.6

Measure #1551

Measure Title: Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Insufficient testing for all-payer cohort**: Panel members noted that there are insufficient testing results to support the non-FFS cohorts.
    - **Developer Response 1:** CORE has decided to change the measure specifications to limit the measure to the over 65, Medicare FFS patient population.
- **Issue 2: Measure specifications, 25-case minimum.** Please explain how the 25-case cut off was established

- **Developer Response 2:** The 25-case minimum was established when these outcome measures were first implemented in public reporting. CMS chose a cutoff of 25 eligible cases to align with the minimum volume requirement used in the publicly reported process of care measures.

- **Issue 3: Low numbers of outliers identified by performance categories:** A Panel member expressed concern about the relatively low proportion of performance outliers.
  - **Developer Response 3:** Please note that performance categories are an implementation issue – CMS chooses to identify outliers based on 95% interval estimates, akin to 95% confidence intervals, which is a conservative approach to identifying performance outliers. The median odds ratio suggests a meaningful increase in the risk of readmission if a patient is admitted with THA/TKA at a higher risk hospital compared to a lower risk hospital. A value of 1.25 indicates that a patient's risk of readmission is 25% greater in a higher-risk hospital than a lower-risk hospital. This variation in rates suggests there are differences in the quality of care received across hospitals performing THA/TKA procedures on Medicare FFS patients.

Validity

- **Issue 1: Model odds ratios.** One panel member asked about the Interpretation of odd ratios, especially for factors associated with lower risk of readmission. Could some of these 'protective' factors be due to collinearity with other risk-factors? Were results assessed for clinical plausibility?
  - **Developer Response 1**: The selection of risk factors for these measures is based on empirical analysis, prior literature, and clinical judgment. The process allows for the selection of variables that are protective in relation to the outcome if they were shown to be significant or of clinical importance. We do test for collinearity with variables that may be clinically similar, and in those instances we often combine similar conditions into one risk variable.
- **Issue 2: Social risk factor adjustment interpretation of results**: One Panel member noted that the empiric data does not support the developer's decision not to include adjustment for social risk factors.
  - **Developer Response 2**: The decision to risk adjust is based both on the empiric results (impact on model and measure scores), the conceptual model (hospitals are better able to mitigate the influence of social risk factors on the measured outcome than clinicians), and CMS' policy decision to adjust for dual eligibility at the program level (within the Hospital Readmission Reduction Program or HRRP). In HRRP, hospitals are stratified into peer groups by the proportion of dual eligible patients and are scored based on their performance within peer groups. Therefore adjustment for social risk factors affects payment to providers, but not the public reporting of the quality measures. It is also consistent with the Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation's (ASPE's) recommendation that quality measures that are used for public reporting should not be risk adjusted (ASPE 2020). As noted in our submission, CMS also confidentially reports disparities in the readmission measures to hospitals so that they have more detailed, actionable information about their patient population's social risk.

Citation:
Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed October 2, 2020.

- **Issue 3: Social risk factor adjustment, net reclassification index.** One Panel member commented that the developers should consider an analysis such as net reclassification index or some other measure to guide the decision of adjusting the measure for social risk.
    - **Developer Response 3**: We agree that model performance should not be judged based on the c-statistic alone. In the context of evaluating hospital performance, it is relevant to assess the impact on hospital performance. We have looked at both the net reclassification index (NRI) and the integrated discrimination improvement (IDI) a few years ago in our development work. We have decided not to use them given some important findings and concerns raised by others.

      Citations:
      Pepe M, Fan J, Feng Z, Gerds T, Hilden H (2015) The Net Reclassification Index (NRI): A misleading measure of prediction improvement even with independent test data sets. Stat Biosci 7:282-295
      Hilder J, Gerds T (2013) A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med 33(19):3405-14.
- **Issue 4: C-statistic**: One panel member noted that the discrimination of the model is weak, suggesting that there are factors relevant to the outcome of interest that are not accounted for in the model.
    - **Developer Response 4**: We ask that the Panel 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 the outcome 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. As noted by another SMP reviewer: "Model discrimination [~0.65] is acceptable for a readmission model; readmission models tend to have lower C stats compared to mortality models because more of the mechanism for readmission lie outside the hospital."
- **Issue 5: Relationship between mortality and readmission:** One Panel member noted that "death is a competing variable for readmission…a site with a high mortality rate may have a lower readmission rate because the patients don't make it back to the hospital."
    - **Developer Response 5**: While this has not yet been established for THA/TKA procedures, in a cohort study of more than 5 million Medicare fee-for-service hospitalizations for heart failure, acute myocardial infarction, and pneumonia from 2008 to 2014, reductions in hospital 30-day readmission rates were weakly but significantly correlated with reductions in 30-day mortality rates after hospital discharge (correlation coefficients, 0.066, 0.067, and 0.108, respectively) (Dharmarajan et al., 2017). This suggests there is no meaningful association between mortality and complications for these conditions. This argues against the hypothesis that increased mortality (or

complications) explains decreased readmissions. Furthermore, CMS has developed and implemented a complication measure that includes mortality as an outcome, so that the relationship between the two metrics can be monitored.

Citation:

Dharmarajan K, Wang Y, Lin Z, Normand ST, Ross JS, Horwitz LI, Desai NR, Suter LG, Drye EE, Bernheim SM, Krumholz HM. Association of Changing Hospital Readmission Rates With Mortality Rates After Hospital Discharge. JAMA. 2017 Jul 18;318(3):270-278.

- **Issue 6: High readmission rates and poor vs. good care quality.** One Panel member noted that "the developers make the presumption that higher readmission reflects poorer quality – therefore a higher readmission rate may in fact be associated with improved care."
    - **Developer Response 6:** In past analyses we have examined the top reasons (diagnosis discharge category or CCS) related to the readmission following a THA/TKA procedure, shown below. The analyses shown below in Table 1 were performed with claims data for Medicare FFS Patients from 2013-2016. These results suggest that more than half of the reasons for return to the hospital were related to a complication from the procedure. This, in addition to feedback from the Technical Expert Panel that established face validity during development, also provides further evidence of validity of the THA/TKA readmission measure as a quality measure.

      Table 1: Discharge diagnosis categories (CCSs) for Medicare FFFS patients readmitted to hospitals within 30 days following THA or TKA procedure (2013-2016).

COUNT	PERCENT (%)	CCS_CATEGORY	CCS_CATEGORY_DESCRIPTION
3584	25	237	Complication of device; implant or graft
2027	14	238	Complications of surgical procedures or medical care
785	5	2	Septicemia (except in labor)
517	4	153	Gastrointestinal hemorrhage
506	4	106	Cardiac dysrhythmias
429	3	226	Fracture of neck of femur (hip)
412	3	108	Congestive heart failure; nonhypertensive
337	2	122	Pneumonia (except that caused by tuberculosis or sexually transmitted disease)
287	2	157	Acute and unspecified renal failure
276	2	145	Intestinal obstruction without hernia
268	2	103	Pulmonary heart disease
266	2	100	Acute myocardial infarction

COUNT	PERCENT (%)	CCS_CATEGORY	CCS_CATEGORY_DESCRIPTION
256	2	197	Skin and subcutaneous tissue infections
242	2	159	Urinary tract infections
229	2	230	Fracture of lower limb
217	2	60	Acute posthemorrhagic anemia
201	1	118	Phlebitis; thrombophlebitis and thromboembolism
193	1	109	Acute cerebrovascular disease
188	1	55	Fluid and electrolyte disorders
165	1	95	Other nervous system disorders
164	1	146	Diverticulosis and diverticulitis
158	1	135	Intestinal infection
131	1	59	Deficiency and other anemia
124	1	127	Chronic obstructive pulmonary disease and bronchiectasis
118	1	149	Biliary tract disease
108	1	155	Other gastrointestinal disorders
108	1	211	Other connective tissue disease
100	1	131	Respiratory failure; insufficiency; arrest (adult)
74	1	117	Other circulatory disease
72	1	204	Other non-traumatic joint disorders

- **Issue 7: Empiric validity.** Panel members expressed concern about the measures chosen as comparator measures for empiric validation. Some panel members wished the developer has used comparators such as process measures that would be expected to be related to the outcome.
    - **Developer Response 7:** CORE met with measure experts to identify candidate comparator measures for validation. Unfortunately there are few available relevant measures and even fewer relevant measures with publicly available data for us to use for such analyses, particularly of the type that would be related to the measure's focus (readmission).

- **Issue 8: Rationale for combining hip and knee procedure outcomes in one measure**: One panel member noted that it was "not clear a priori why it is appropriate to mix hip and knee replacement surgery. Is there no difference in outcome between these two operations?"
    - **Developer Response 8**: We considered whether to develop separate measures for patients undergoing THA or TKA procedures or to combine patients undergoing either procedure into a single hospital quality measure. To inform that decision we consulted with the working group and conducted analyses to examine the average length of stay, and mortality, complication, and readmission rates for each procedure. Based on those analyses (see results in the original methodology report) and in consultation with the working group, we combined these patient cohorts for the readmission measure for several reasons, including:
        - A large proportion of THA and TKA procedures are elective and performed in similar patient cohorts for similar indications (e.g. osteoarthritis).
        - The rates and types of complications were similar
        - The mortality and readmission rates are similar
        - Hospitals develop protocols/programs for lower extremity total joint arthroplasty, rather than THA and TKA separately
        - Combining admissions for both procedures will provide greater power to detect hospital-level variation to enable quality improvement

Measure #3599

Measure Title: Pediatric Asthma Emergency Department Use

Measure Developer/Steward: University of California San Francisco

---

Validity

- **Issue 1:** Exclusions reported were tested using only data from MA.

  This issue was raised during the last review of this measure, and developers responded with additional information on exclusion rates for CA (which were very small). Please include this additional information in this submission!
    - **Developer Response 1:** Our apologies for this oversight. We assessed missingness in the 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.
- **Issue 2:** Concern related to numerator in section 2 above; is what is being measured truly related to asthma management?
    - **Developer Response 2:** There is a robust body of literature supporting the importance of asthma management to reduce acute utilization (emergency department or hospitalization). This is summarized in NHLBI's NAELPP guideline "Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma-Full Report 2007". Specific examples include: Schatz and colleagues study describing the relationship between asthma control and asthma exacerbations in managed care (1), and Fuhlbrigge et al's

confirmation that medications can work to reduce ED visits for asthma but are used sub optimally (2). When children with asthma experience adequate management of chronic conditions and have access to coordinated care, a reduction in hospital rates is likely to occur. (3) Children who are linked to continuous care utilize less overall care, including ED care. (3)

(1) Schatz M, Zeiger RS, Yang ST, et al. "Relationship of asthma control to asthma exacerbations using surrogate markers within a managed care database." Am J Managed Care. 16(5):327-333, 2010.

(2) Fuhlbrigge A, Carey VJ, Adams RJ. "Evaluation of asthma prescription measures and health system performance based on emergency department utilization." Med Care 42(5):1-7, 2004.

(3) Cooley W, McAllister J, Sherrieb K, Kuhlthau K. *Improved outcomes associated with medical home implementation in pediatric primary care.* Pediatrics, 2009. **124**(1): p. 358-364.

In addition, we provide evidence of the relationship between the measure (pediatric asthma ED visits) and asthma management, in our report of the VCHIP collaborative success in decreasing pediatric asthma utilization through improved NHLBI recommended clinical care processes in practices participating in a 7-month learning collaborative. This is described in 2b1.2 in the testing attachment.

- **Issue 3:** Concerns with the low R-squared values for the CA plans (13%), which is very different value from the R-squared value for the MA plans (56%). Could it be a case that the model is a better fit for APCD than Medicaid only? The Medicaid only population may have less variation in the community risk factors (% below poverty, education, unemployment) which may make the models less predictive.
    - **Developer Response 3:**
      We assessed the variation in community risk factors in CA and MA, to assess this hypothesis. We found the following:

Metric	MA mean	MA SE	CA mean	CA SE	Ratio of CA SE/MA SE
Median household income	66269	34.61233	54368.44	11.01509	0.318
% housing public assistance	17.01975	0.015342	5.908931	0.0021989	0.143
% households below FPL	14.54965	0.012076	17.40438	0.0054135	0.448
Number of households	10124.46	6.602852	14183.93	3.517135	0.533
% female headed household	15.71511	0.009999	17.55645	0.0032822	0.328
% unemployed	8.717494	0.004627	6.450902	0.0011429	0.247

***SE: standard error**

- The ratio of CA standard error to MA standard error is consistently less than one, supporting the hypothesis that the Medicaid population in CA has less variation in

community risk factors compared to the MA population. These findings support the hypothesis from the reviewer that the difference between the two states in their R-squared values is due to less variation across Medicaid members in CA compared to MA. This explains why the R squared values are dissimilar, with less of the variation in asthma ED visits explained by the risk-adjustment variables in CA compared to variation in asthma ED visits explained by risk adjustment variables in MA. These results should be taken in context with the rest of the validity data presented. While it would be nice to see the same R-squared across both states, it is not a fatal flaw in the measure, since R-squared is only one metric for judging validity.

- The rationale behind testing the measure in the CA dataset is to assess whether the model is overly specified for MA data. In order to test whether we had over-specified the risk adjustment model using MA data, we built the model afresh using CA data and the same backwards selection process. We started out with the same set of patient variables as in the base model (avg age, gender, chronic condition indicators) and the available community social determinants of health risk factors in the CA data (Median household income, % housing public assistance, % households below FPL, number of households, % female headed household,% unemployed). While the R-squared value went as high as 0.20 while including almost all risk factors in CA, the R-squared for the equivalent model in MA also improved similarly. This supports the idea that the difference in R-squared is not due to over-specification in the models, but rather that it reflects a more homogenous population in CA Medicaid alone vs. the members in the APCD data.
- Of note, we decided not to use the expanded set of variables from the CA backward selection process, since the original risk adjustment model reflected an evidence-based set of variables (percent of households below the poverty level, percent of adults over 25 with less than a high school education, percent male unemployment for 25 to 60 year olds). These specific variables have been validated as measures of SES using factor analysis and found to be associated with increased readmissions as well as direct measures of allostatic load or physiological stress (Martsolf, 2016 and Bird, 2010).

- **Issue 4:** Construct validity: assessed correlation between performance on measure and performance on related and unrelated HEDIS measures. There was a concern from one reviewer that there was low correlation with HEDIS measures.
  - **Developer Response 4:**
    While the correlations with the related measures are not strong correlations, we would not expect them to be strongly correlated, for several reasons. 1) Multiple factors go into Pediatric Asthma ED Use (primary care access, successful medication acquisition and adherence, access to good air quality and dust-free housing); 2) HEDIS measures of asthma medication use include adults as well as pediatric patients, 3) Immunizations for children <2 years old, which was more strongly correlated with asthma ED utilization (0.33, P=0.02), require different efforts than optimizing care for pediatric asthma patients. The effect sizes for all the related measures in the Table are larger than the unrelated adult measures, indicating the construct validity of Pediatric Asthma ED Use. See table in testing attachment in 2b1.3 to review original testing results.

- **Issue 5:** Concerns remain with the number of plans that have a relatively high percentage of members who are missing social risk factor data. The social risk factors are key adjustment variables in the risk adjustment models. 20,000 of 85,000 members are in plans for which 10% of more of members are missing social risk data.
    - **Developer Response 5:**

      It is helpful to note a few things to place this in context. See table of missingness for MA in the original testing attachment question 2b6.2. There are 8 plans with 10% or more members are missing data; most of these plans have very few members. The two exceptions are 8647 with 4,220 members and 291 with 13,247, taken together accounting for 88% of members in plans with 10% or more members missing. However, for both these plans, there is still substantial data available, with 90% of members with data on SES for Plan 8647 and 76% of members with data on SES for Plan 291.
      The total number of people with missing data is small (5935 people), which is 6.9% of 85K members.
      Also of note, CA has very little missing data (see above).
- **Issue 6: "**Given that secondary diagnosis for asthma may be unrelated to the reason for ER visit/admission, the validity of the measure as constituted has not been established."
    - **Developer Response 6:** We addressed this concern in a peer-reviewed paper using this measure. One clarification: the measure only includes asthma diagnoses in the first or second diagnostic spot. This is different from ANY secondary diagnosis (meaning, any diagnosis other than the first, primary diagnosis). We included claims with a second diagnosis of asthma because the primary diagnosis was often a related symptom (e.g., fever, wheezing) or a known asthma trigger (e.g., upper respiratory tract infection, pneumonia, influenza). We checked this assumption by tabulating the primary diagnoses for all the ED visits that were included with asthma in the second diagnostic spot, to confirm this. In addition, we performed sensitivity analyses to assess whether the primary findings of the paper changed if we only included ED visits with a primary diagnosis of asthma. The paper assessed the relationship between pediatric asthma ED utilization and diagnoses of anxiety and/or depression. We present the work from that paper below. (citation: Bardach et al. Depression, Anxiety, and Emergency Department Use for Asthma. *Pediatrics* 2019).
    - The sensitivity analysis demonstrated two main findings—that relationships were similar using both definitions (primary analysis: ED visits with asthma in 1st or 2nd spot (Table 2), and sensitivity analysis: ED visits with asthma diagnosis in 1st spot only (eTable8 below); that dropping visits with asthma diagnoses in the second spot led to a loss of almost half of the numerator events. The first finding supports the validity of inclusion of the visits using asthma in the 1st of 2nd spot. The second finding supports the decision to keep the more liberal definition, in order to avoid losing half the events, which would limit the utility of the measure to statistically differentiate between health plans.

      Methods:
      Sensitivity analysis: "Because numerator events were identified using asthma as the first or second diagnosis, we separately re-ran the analyses only using ED visits with asthma as a first diagnosis in the numerator."

Results:

PRIMARY ANALYSIS

Table 2. Asthma-related ED visits rate per 100 child-years by patient characteristics

	Asthma-related ED visits/100 child-years, Rate (unadjusted, 95% Confidence Interval)	**Relative Rate of Asthma-related ED visits/100 child-years (adjusted, 95% Confidence Interval)**	***P*-value***
Mental health conditions			
No anxiety or depression	15.2 (14.1-16.3)	Reference	Reference
Anxiety only	18.6 (16.6-20.6)	1.22 (1.10-1.35)	< 0.001
Depression only	24.8 (20.7-28.8)	1.43 (1.23-1.62)	< 0.001
Anxiety and depression	30.5 (27.5-33.5)	1.80 (1.60-2.00)	< 0.001

SENSITIVITY ANALYSIS

eTable 8. New numerator definition: ED visits for asthma as the first diagnosis only				
	Rate	95% CI Lower Limit	95% CI Upper Limit	P-value*
No Anxiety or Depression	7.7	7.2	8.3	Reference
Anxiety Only	8.5	7.3	9.7	0.203
Depression Only	9.5	7.4	11.6	0.053
Anxiety and Depression	11.0	9.3	12.7	<0.001

* p-value for results of multivariate comparison testing model, adjusted for age category, gender, insurance type, and chronic disease status. CI: Confidence Interval.

Subgroup 3

Measure #1893

Measure Title: Hospital, 30-day, all-cause, risk-standardized mortality rate (RSMR) following chronic obstructive pulmonary disease (COPD) hospitalization

Measure Developer/Steward: Yale CORE/CMS

Reliability

- **Issue 1: Insufficient testing for all-payer cohort**: Panel members noted that there are insufficient testing results to support the all-payer cohorts.
    - **Developer Response 1:** CORE has decided to change the measure specifications to limit the measure to the over 65, Medicare FFS patient population.

- **Issue 2: Specifications; definition of exclusions:** Panel members asked the developer for clarification regarding exclusions. One panel member asked specifically for clarification regarding exclusion of claims records with sex other than "male" or "female" due to concern that vulnerable patients may be negatively affected.
    - **Developer Response 2:** As stated in S8 of the submission/ITS form, the measure excludes index admissions for patients:
    - **With inconsistent or unknown vital status or other unreliable demographic (age and gender) data**. As stated in the submission form, inconsistent vital status or unreliable data are identified if any of the following conditions are met 1) the patient's age is greater than 115 years: 2) if the discharge date for a hospitalization is before the admission date; 3) if the patient has a sex other than 'male' or 'female'. The rationale for this exclusion is that reliable and consistent data are necessary for valid calculation of the measure. Regarding the concern about vulnerable populations we agree with the reviewer that this is an important concern. It is an evolving area and CORE is working with CMS to better understand if there are any unintended consequences regarding the way sex is defined and collected in claims data. We do not have this information on hand to share on short notice but we will be calculating the proportion of claims that are affected by this exclusion.
    - **Enrolled in the Medicare hospice program or used VA hospice services any time in the 12 months prior to the index admission**, including the first day of the index admission. Here we clarify how these admissions are defined: the Medicare Enrollment Database (EDB) was used to identify patients who were enrolled in hospice.
    - **Discharged against medical advice (AMA).** As stated in the submission/ITS form, discharges against medical advice (AMA) are identified using the discharge disposition indicator in claims data.

Validity

- **Issue 1: Low numbers of outliers identified by performance categories:** A Panel member expressed concern about the relatively low proportion of performance outliers.
    - **Developer Response 1:** Please note that performance categories are an implementation issue – CMS chooses to identify outliers based on 95% interval estimates, akin to 95% confidence intervals, which is a conservative approach to identifying performance outliers. This implementation approach not part of the NQF measure specifications. As noted by other SMP reviewers, the median odds ratio of 1.26 is clinically significant. The median odds ratio suggests a meaningful increase in the risk of mortality if a patient is admitted with COPD at a higher risk hospital compared to a lower risk hospital. A value of 1.26 indicates that a patient's risk of mortality is 26% greater in a higher risk hospital than a lower risk hospital. This variation suggests there are differences in the quality of care received across hospitals for COPD. This evidence supports continued measurement to reduce the variation.

Measure #0141

Measure Title: Patient Fall Rate

Measure Developer/Steward: American Nurses Association (ANA)

Reliability

- **Issue 1:** Clarification of monthly/quarterly rate
    - **Developer Response 1:** Hospitals submit monthly data on patient fall events (numerator) and patient volume (denominator). The current quality reporting timeline for NDNQI is quarterly to reduce data submission burden for hospitals and includes quarterly rates.

Validity

- **Issue 1:** Exclusion criteria was unclear and inconsistent on MIF and testing forms
    - **Developer Response 1:** There are no exclusion criteria based on patient or event-related factors (e.g. patient age, diagnosis, refusal of test/treatments). All patient fall events occurring on patient care units are included. The current measure testing information does not include all unit types found in inpatient settings but comprises the most common and standardized unit type classifications.
- **Issue 2:** Clarification on missing data
    - **Developer Response 2:** In order to receive benchmarking reports, hospitals must submit complete data, and error reports and data validations are triggered if organizations fail to submit complete data and the data are not included in the benchmarking database.
- **Issue 3:** Concerns over lack of risk adjustment
    - **Developer Response 3:** The measures are not risk-adjusted, but the hospital level measure is a weighted standardized score that adjusts for the differential risk of patients across care settings (i.e. unit types) and weighted for the volume of patients in each unit type. This approach does account for the mix of nursing unit types (as suggested by Panel Member #9). The concerns of the committee over patient-risk factors are warranted, however, the application of this measure is not intended as provider performance or payment programs, but as organizational quality and patient safety. As such, the measure is not designed to capture patient-level information beyond age, and gender, and risk assessment scores. Further, the measure only captures fall events, and does not include any information on patients who did not experience a fall on the unit. The stratification across unit types provides some adjustment for case mix relative to fall risk (e.g. falls in critical care units are more rare than in medical units where patients are more ambulatory), but a full risk-adjusted model is not possible with these data or this measure. However, there are currently no other comprehensive patient falls measure for inpatient settings, and it remains one of the most common adverse patient events. Given the current measure gap and the importance of the measure, the unit type stratification methods to adjust for patient mix are adequate to meet the intended use of the measure.
- **Issue 4:** Method of validity testing – concerns over measures used in convergent validity testing
    - **Developer Response 4:** The measures used in the convergent validity testing are the best available, given the unique nature of NDNQI data. Other safety measures are focused solely on hospital-level performance. NDNQI uses nursing-unit level, which

provides an important perspective for nursing quality and patient safety but limits the availability of similar data.

Other General Comments

- **Developer comment:** The risk-adjustment concerns of the SMP are noted. However, there are currently no other comprehensive patient falls measure for inpatient settings, and it remains one of the most common adverse patient events. Given the current measure gap and the importance of the measure, the unit type stratification methods to adjust for patient mix are adequate to meet the intended use of the measure. Although standard rules for evaluation the scientific acceptability of measures is necessary (and appreciated), the intended use of the measures should be taken into account. If payment models were attached, where undue penalties might further jeopardize the safety of patients, a fully risk-adjusted measure would be more appropriate. However, every patient fall is a serious safety event, regardless of the underlying patient characteristics.

Measure #0202

Measure Title: Falls with Injury

Measure Developer/Steward: American Nurses Association (ANA)

Reliability

- **Issue 1:** Clarification of monthly/quarterly rate
    - **Developer Response 1:** Hospitals submit monthly data on patient fall events (numerator) and patient volume (denominator). The current quality reporting timeline for NDNQI is quarterly to reduce data submission burden for hospitals and includes quarterly rates.

Validity

- **Issue 1: Exclusion criteria was unclear and inconsistent on MIF and testing forms**
    - **Developer Response 1:** There are no exclusion criteria based on patient or event-related factors (e.g. patient age, diagnosis, refusal of test/treatments). All patient fall events occurring on patient care units are included. The current measure testing information does not include all unit types found in inpatient settings but comprises the most common and standardized unit type classifications.
- **Issue 2: Clarification on missing data**
    - **Developer Response 2:** In order to receive benchmarking reports, hospitals must submit complete data, and error reports and data validations are triggered if organizations fail to submit complete data and the data are not included in the benchmarking database.
- **Issue 3: Concerns over lack of risk adjustment**
    - **Developer Response 3:** The measures are not risk-adjusted, but the hospital level measure is a weighted standardized score that adjusts for the differential risk of patients across care settings (i.e. unit types) and weighted for the volume of patients in each unit type. This approach does account for the mix of nursing unit types (as suggested by

Panel Member #9). The concerns of the committee over patient-risk factors are warranted, however, the application of this measure is not intended as provider performance or payment programs, but as organizational quality and patient safety. As such, the measure is not designed to capture patient-level information beyond age, and gender, and risk assessment scores. Further, the measure only captures fall events, and does not include any information on patients who did not experience a fall on the unit. The stratification across unit types provides some adjustment for case mix relative to fall risk (e.g. falls in critical care units are more rare than in medical units where patients are more ambulatory), but a full risk-adjusted model is not possible with these data or this measure. However, there are currently no other comprehensive patient falls measure for inpatient settings, and it remains one of the most common adverse patient events. Given the current measure gap and the importance of the measure, the unit type stratification methods to adjust for patient mix are adequate to meet the intended use of the measure.

- **Issue 4: Method of validity testing – concerns over measures used in convergent validity testing**
    - **Developer Response 4:** The measures used in the convergent validity testing are the best available, given the unique nature of NDNQI data. Other safety measures are focused solely on hospital-level performance. NDNQI uses nursing-unit level, which provides an important perspective for nursing quality and patient safety but limits the availability of similar data.

Other General Comments

- **Developer comment:** The risk-adjustment concerns of the SMP are noted. However, there are currently no other comprehensive patient falls measure for inpatient settings, and it remains one of the most common adverse patient events. Given the current measure gap and the importance of the measure, the unit type stratification methods to adjust for patient mix are adequate to meet the intended use of the measure. Although standard rules for evaluation the scientific acceptability of measures is necessary (and appreciated), the intended use of the measures should be taken into account. If payment models were attached, where undue penalties might further jeopardize the safety of patients, a fully risk-adjusted measure would be more appropriate. However, every patient fall is a serious safety event, regardless of the underlying patient characteristics.

Measure #3592

Measure Title: Global Malnutrition Composite Score

Measure Developer/Steward: Avalere Health/Academy of Nutrition and Dietetics

---

Reliability

- **Issue 1:** Component Measure #1 Specification
    - **Developer Response 1:** The measure specification is not “those patients who were screened within 48 hours prior to admission how many patients were identified as at-

risk through that screening" as suggested by the Panel Member. The measure is all admitted patients who have a screening result no earlier than 48 hours prior to admission but has not cap on when AFTER admission that screening may occur. In the hospital, most patients are screened within 24 hours of admission, some hospitals implement screening even prior to admission or during an observation stay. These are documented in the same standard place regardless of the admission type. The inclusion criteria for this measure are patients who have a hospital admission, and THEN a screening encounter that occurred no earlier than 48 hours prior to the admission. Therefore, all patients screened at any point are admitted.

- **Issue 2:** Combination of four measures into composite score

    - **Developer Response 2:** As shown in the item-level analyses, the strongest associations with outcomes of 30-day readmissions and length of stay are those elements of a malnutrition diagnosis and the nutrition care plan. Not all patients who are found at-risk are fully malnourished and necessitate a diagnosis and care plan. However, the entire care workflow begins at admission with first identification of risk leading to those who need a care plan if diagnosed. The measure is reliable and valid because with the current flat average and a case minimum, the signal to noise analysis resulted in an above average ICC.

- **Issue 3:** Nesting of sites within health systems
    - **Developer Response 3:** The composite measure is intended to be implemented at the site level. However, since data are reported at the site-level, we do not have data to differentiate between individual providers. As a result, direct derivation of the between- and within-site provider variance components was not directly supported by the raw data, so we elected to utilize a system-level ICC as a surrogate, understanding that it would likely underestimate the true site-level ICC.
    - To address the reviewer's concerns, however, we developed a complementary approach to ICC estimation whose results are applicable at the site level and continue to support the conclusion that the composite measure's reliability falls within the acceptable range. Specifically, prior to fitting the data to the mixed effects model, five (5) versions of the patient-level data set were generated by randomly assigning patients to provider "blocks" of sizes 10, 20, 30, 40, and 50, respectively, within each practice site. Pragmatically, the provider blocks can be regarded as idealized practice sizes that 1) might be encountered if patients were randomly paneled to providers and 2) the number of providers practicing at each site was held constant across health systems.
    - After creating the provider blocks, the intercept-only mixed effects model was then refitted to each data set, incorporating practice site—instead of health system—as the random effect. In turn, ICC's were generated using the between- and within-site variance components extracted from each model. The model-derived ICC's are presented in the following table:

Number of Providers	Mean Panel Size	ICC

10	320	0.89
20	160	0.82
30	106	0.76
40	80	0.73
50	64	0.65

    - As before, the results suggest that composite measure reliability is satisfactory and falls within the acceptable range specified by NQF. Moreover, we anticipate that the between-provider variance and, therefore, the site-level ICC will be higher in the real world, reflecting natural variation that arises from fluctuations in the number of providers per site and panel sizes.
- **Issue 4:** Minimum number of individual performance measures required for scoring
    - **Developer Response 4:** The inconsistency noted between the specifications and the testing documents will be corrected in the updated submission, the minimum should be three measures. The measures are indeed conditional, a hospital would not qualify for measures 3 or 4 for instance if there are not sufficient cases in measure 2.
- **Issue 5:** States reflected in testing
    - **Developer Response 5:** The testing was completed with the individual hospitals nested within their integrated delivery networks or health systems; this was not an analysis at the state level as proposed by Panel Member 8. Health system is the actual health system of hospitals not the state in which a health system was located in.
- **Issue 6:** Ranges of ICC suggesting ICC test model 1 was poor to marginal.
    - **Developer Response 6:** According to multiple sources and studies, an ICC score above 0.5 indicates at least moderate reliability (certainly not poor), and a score above 0.75 is good reliability with above 0.9 indicating excellent. One reviewer suggested that the test without the case minimums suggested that the reliability was “poor”. That statement is not supported by any evidence we have encountered on ICC ranges. One example scale is provided here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4913118/

Validity

- **Issue 1:** Differentiation of Groups
    - **Developer Response 1:** The distribution demonstrated that while a large portion of the hospitals were high performing (they have been implementing quality improvement on malnutrition for more than 4 years at this point), almost half of the hospitals (40%) performed below the middle tier of performance indicating substantial room for improvement. Once hospitals improve on the core steps necessary to provide the appropriate care, they can quickly begin to improve on their performance scores.
- **Issue 2:** Confidence Intervals

- **Developer Response 2:** We will review the use of confidence intervals at the hospital level for proper performance tier placement and make refinements in advance of the meeting and final submission deadline.

Other General Comments

- **Issue 1:** Method for Establishing Validity (end of Page 8, Panel Member 3)
  - **Developer Response 1:** Please review tables 3, 4, and 5 which outline the results of the univariate analysis which was conducted to understand the association between critical components of the composite with known outcomes associated with malnutrition. All major components: malnutrition screening, assessment, malnutrition diagnosis and care plan indicate that patients who require these are more likely to require longer treatment (longer LOS) and be at higher risk for readmission. This univariate analysis does not intend to measure whether have a malnutrition diagnosis for example is associated with a *decreased* risk of either longer LOS or 30-day readmission. It is to demonstrate the strong correlation and directionality between these individual variables and associated outcomes. The score level analysis was to demonstrate that the conceptual result of the composite score (malnutrition diagnosis AND a nutrition care plan for those who were at-risk) resulted in a differential outcome than that of the case where there is a malnutrition diagnosis but not a nutrition care plan. The increased LOS outcome is more likely an indication of additional time spent in the hospital to develop and implement a nutrition care plan than a causal relationship.

Measure #3235

Measure Title: Hospice and Palliative Care Composite Process Measure

Measure Developer/Steward: Abt Associates/CMS

Reliability

- **Issue 1:** A review asked for a clarification on the interpretation of the intraclass correlation coefficient (ICC) score we calculate to support reliability.
  - **Developer Response 1:** The reviewer was correct in their assumption that this value was the correlation coefficient between two measure scores, each calculated from a randomly selected half of the episodes from each provider. We are pleased that the reviewer believed the reliability score to be high in that case.
- **Issue 2:** Reviewers suggested the “signal-to-noise” reliability testing should be removed
  - **Developer Response 2:** Thank you for this comment. Our reliability testing (including signal-to-noise analysis) is an update of the approach used in the original endorsement submission. If the panel would like us to exclude signal-to-noise testing, and if the ICC results alone are sufficient, then we could do so in the final submission (and focus our reliability testing on ICC and stability results), otherwise we can retain it.

Validity

- **Issue 1:** A reviewer expressed a concern with Table 2 (regarding the exclusion of individuals under the age of 18), results did not add up to 100% and therefore there could have been a computational mistake
    - **Developer Response 1:** We appreciate the reviewer's keen eye and brining this to our attention. We reviewed the materials and have concluded that the numbers were correct, but the tables could be better labeled. In fact, the numbers were not supposed to add up to 100%, but rather, indicate two slightly different groups; i.e., (#1) the percent of pediatric patients in the numerator and, separately, (#2) the percent of non-pediatric patients in the numerator. This might be clearer in the accompanying text: "*From the results in this table, we see that non-pediatric patients are only slightly more likely to receive the desired care process compared with pediatric patients (for example, 96.88 percent received a pain screening within 2 days of admission, compared with 98.24 percent among patients who were at least 18 years old)*". We propose to update the labeling in Table 2 of our final submission to better clarify the intent of the table. Our proposed changes are: "Percentage of pediatric patients in the numerator" to be relabeled as "Percent of pediatric patients that received the desired care process" and "Percentage of non-pediatric patients in the numerator" to be relabeled as "Percent of non-pediatric patients that received the desired care process".
- **Issue 2:** Concerns were raised about the approach to validity, by correlating the composite with its individual (NQF-endorsed) component measures. Reviewers suggested that we validate against other, independent measures.
    - **Developer Response 2:** At the time of original endorsement submission, the individual components were the only other quality measures available for validation. Since then, the program has added measures calculated from CAHPS survey (NQF #2651). Unfortunately, there is a misalignment between the two in timing and moreover conceptually: the measure for which we're seeking endorsement (#3235) covers processes at admission, whereas the #2651 is an outcome measure covering experience of care over the length of the episode. That said, we will empirically explore correlations between #2651 and #3235 as well as discuss its use further internally. If appropriate, we will include these findings in the final submission.
- **Issue 3:** One reviewer questioned the exclusion of patients under the age of 18.
    - **Developer Response 3:** The reason for this exclusion was that the NQF-endorsed component of this composite similarly excluded pediatric hospice patients from their denominators. Therefore it would not be appropriate to include these individuals in the composite. We note another reviewer agreed this exclusion was appropriate.
- **Issue 4:** One reviewer questioned the omission of risk adjustment from this measure.
    - **Developer Response 4:** Process measures such as these are typically not risk adjusted and so risk adjustment was not included in the original or maintenance of endorsement submission. We note another reviewer agreed it was appropriate no risk adjustment was conducted for this measure.
- **Issue 5:** One reviewer questioned a reference missing data to establish validity (in 27b)
    - **Developer Response 5:** We thank the reviewer for this. We will review and add more language as appropriate. This analysis was simply an update of the same analysis used in

the original submission, and seeks to statistically describe the rate of various items missing. Our conclusions were that missing data is overall a low prevalence, and thereby we are not concerned as to the accuracy of our results, via data completion. We will review our argument and if the point could be removed we will do so for the final submission.