NATIONAL QUALITY FORUM
Driving measurable health improvements together

Scientific Methods Panel Discussion Guide

SPRING 2021 EVALUATION CYCLE
March 30-31, 2021

This guide is funded by the Centers for Medicare & Medicaid Services under contract HHSM-500-2017-00060I - HHSM-500-T0001.

Contents

Background

The Scientific Methods Panel (SMP) provides the National Quality Forum (NQF) Standing Committees with evaluations of the scientific acceptability of submitted complex measures, specifically, the "must-pass" sub-criteria of reliability and validity, using NQF's standard measure evaluation criteria for both new measures and maintenance measures.

This discussion guide contains details of the complex measures submitted for evaluation during the spring 2021 measure evaluation cycle. It also contains summaries of the SMP's preliminary measure analyses and responses to these analyses composed by measure developers. The SMP utilizes this document during their measure evaluation meetings to facilitate conversations between the SMP members, measure developers, and NQF staff.

During this cycle, the SMP evaluated 29 complex measures. After the SMP's preliminary review and vote, measure developers were provided the opportunity to respond to the SMP's feedback. Measures were slated for discussion and revote at the SMP's measure evaluation meeting if consensus was not reached during the preliminary review, or if a measure did not pass a sub-criterion and the developer organization provided a written response to the SMP's comments. Additionally, SMP members and NQF staff may pull a measure for further discussion as they see fit, even if the measure passed preliminary review. Eight measures are currently up for discussion and revote. Five measures have been pulled by SMP members or NQF staff for further discussion due to concerns raised during the preliminary review, although they have passed NQF's scientific acceptability criterion. Measures not pulled for discussion will not be discussed during the SMP's evaluation meeting.

Following the SMP's review of the complex measures, those that pass on scientific acceptability move on to their respective Standing Committees for a full evaluation. The Standing Committees review the remaining NQF standard measure evaluation criteria (Importance to Measure and Report, Feasibility, Usability and Use, and requirements for Related and Competing Measures), the SMP's feedback, and the scientific acceptability ratings, with the option to either accept or overturn the results. Measures that do not pass the SMP can be pulled by a Standing Committee member for further discussion and revote if it is an eligible measure. A measure is eligible for revote if the SMP found none of the following:

- Inappropriate methodology or testing approach applied to demonstrate reliability or validity
- Incorrect calculations or formulas used for testing
- Description of testing approach, results, or data is insufficient for SMP to apply the scientific acceptability sub-criteria
- Appropriate levels of testing not provided or otherwise did not meet NQF's minimum evaluation requirements

Measures that do not pass the SMP's review and are not pulled for discussion by the Standing Committee can be deferred, revised, and resubmitted for reconsideration in a future cycle. Please refer to *Scientific Methods Panel: Frequently Asked Questions* in NQF's standard measure evaluation criteria for details on this process.

Measures for Discussion: Preliminary Voting Results

Legend: High (H), Moderate (M), Low (L), and Insufficient (I)

Subgroup 1

- #2880 Excess Days in Acute Care (EDAC) After Hospitalization for Heart Failure (HF) (Yale Center for Outcomes Research and Evaluation (CORE)/Centers for Medicare & Medicaid Services (CMS))
    - **Reliability:** H-0; M-8; L-1; I-0 **Pass**
    - **Validity:** H-1; M-3; L-2; I-3 **Consensus Not Reached (CNR)**
- #2881 Excess Days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI) (Yale CORE/CMS)
    - **Reliability:** H-0; M-4; L-5; I-0 **CNR**
    - **Validity:** H-0; M-3; L-3; I-3 **No Pass**
- #2882 Excess Days in Acute Care (EDAC) After Hospitalization for Pneumonia (Yale CORE/CMS)
    - **Reliability:** H-1; M-8; L-0; I-0 **Pass**
    - **Validity:** H-0; M-3; L-3; I-3 **No Pass**
- #3188 30-Day Unplanned Readmissions for Cancer Patients (Alliance of Dedicated Cancer Centers)
    - **Reliability:** H-0; M-7; L-2; I-0 **Pass**
    - **Validity:** H-0; M-3; L-4; I-2 **No Pass**
- #3612 Risk-Standardized Acute Cardiovascular-Related Hospital Admission Rates for Patients With Heart Failure Under the Merit-Based Incentive Payment System (Yale CORE/CMS)
    - **Reliability:** H-0; M-5; L-4; I-0 **CNR**
    - **Validity:** H-1; M-4; L-2; I-2 **CNR**
- #3615 Unsafe Opioid Prescriptions at the Prescriber Group Level (University of Michigan Kidney Epidemiology and Cost Center (UM-KECC))
    - **Reliability:** H-6; M-1; L-1; I-1 **Pass**
    - **Validity:** H-2; M-4; L-1; I-2 **Pass**
- #3616 Unsafe Opioid Prescriptions at the Dialysis Practitioner Group Level (UM-KECC)
    - **Reliability:** H-1; M-6; L-1; I-1 **Pass**
    - **Validity:** H-1; M-5; L-1; I-2 **Pass**
- #3622 National Core Indicators for Intellectual and Developmental Disabilities (ID/DD) Home and Community-Based Services (HCBS) Measures (Human Services Research Institute)
    - **Reliability:** H-3; M-3; L-2; I-1 **Pass**
    - **Validity:** H-0; M-2; L-3; I-4 **No Pass**

Subgroup 2

- #3614 Hospitalization After Release With Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke) (Armstrong Institute for Patient Safety and Quality at Johns Hopkins University)
    - **Reliability:** H-0; M-5; L-1; I-2 **Pass**
    - **Validity:** H-2; M-2; L-3; I-1 **CNR**

Subgroup 3

- #0500 Severe Sepsis and Septic Shock: Management Bundle (Henry Ford Hospital)
  - **Reliability:** H-5; M-1; L-0; I-2 **Pass**
  - **Validity:** H-3; M-2; L-1; I-2 **Pass**
  - **Composite Construction:** H-2; M-3; L-0; I-1 **Pass**
- #0674 Percent of Residents Experiencing One or More Falls With Major Injury (Long Stay) (RTI International/CMS)
  - **Reliability:** H-0; M-6; L-2; I-0 **Pass**
  - **Validity:** H-1; M-6; L-1; I-0 **Pass**
- #0679 Percent of High-Risk Residents With Pressure Ulcers (Long Stay) (RTI International/CMS)
  - **Reliability:** H-0; M-6; L-2; I-0 **Pass**
  - **Validity:** H-2; M-4; L-2; I-0 **Pass**
- #3621 Composite Weighted Average for 3 CT Exam Types: Overall Percent of CT exams for Which Dose Length Product Is at or Below the Size-Specific Diagnostic Reference Level (for CT Abdomen-Pelvis With Contrast/Single Phase Scan, CT Chest Without Contrast/Single (American College of Radiology)
  - **Reliability:** H-5; M-2; L-0; I-1 **Pass**
  - **Validity:** H-0; M-4; L-0; I-4 **CNR**
  - **Composite Construction:** H-2; M-3; L-0; I-1 **Pass**

Measures That Passed (Not Pulled for Discussion): Preliminary Voting Results

Subgroup 1

- #2860 30-Day, All-Cause, Unplanned Readmission Following Psychiatric Hospitalization in an Inpatient Psychiatric Facility (IPF) (Yale CORE/CMS)
  - **Reliability:** H-1; M-8; L-0; I-0 **Pass**
  - **Validity:** H-1; M-6; L-1; I-1 **Pass**

Subgroup 2

- #1598 Total Resource Use Population-Based PMPM Index (HealthPartners)
  - **Reliability:** H-4; M-3; L-0; I-2 **Pass**
  - **Validity:** H-4; M-2; L-1; I-2 **Pass**
- #1604 Total Cost of Care Population-Based PMPM Index (HealthPartners)
  - **Reliability:** H-4; M-3; L-1; I-1 **Pass**
  - **Validity:** H-3; M-4; L-2; I-0 **Pass**
- #2431 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Acute Myocardial Infarction (AMI) (Yale CORE/CMS)
  - **Reliability:** H-3; M-5; L-0; I-0 **Pass**
  - **Validity:** H-1; M-5; L-2; I-0 **Pass**
- #2436 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Heart Failure (HF) (Yale CORE/CMS)
  - **Reliability:** H-5; M-3; L-0; I-0 **Pass**
  - **Validity:** H-2; M-4; L-2; I-0 **Pass**

- #2579 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Pneumonia (PN) (Yale CORE/CMS)
    - **Reliability:** H-5; M-3; L-0; I-0 **Pass**
    - **Validity:** H-2; M-4; L-2; I-0 **Pass**
- #3610 30-Day Risk-Standardized Morbidity and Mortality Composite Following Transcatheter Aortic Valve Replacement (TAVR) (American College of Radiology)
    - **Reliability:** H-0; M-7; L-1; I-0 **Pass**
    - **Validity:** H-3; M-5; L-0; I-0 **Pass**
    - **Composite Construction:** H-3; M-3; L-1; I-1 **Pass**
- #3623 Elective Primary Hip Arthroplasty (Acumen, LLC/CMS)
    - **Reliability:** H-7; M-1; L-0; I-0 **Pass**
    - **Validity:** H-0; M-5; L-2; I-0 **Pass**
- #3625 Non-Emergent Coronary Artery Bypass Graft (CABG) (Acumen, LLC/CMS)
    - **Reliability:** H-4; M-4; L-0; I-0 **Pass**
    - **Validity:** H-0; M-5; L-3; I-0 **Pass**
- #3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels (Acumen, LLC/CMS)
    - **Reliability:** H-4; M-4; L-0; I-0 **Pass**
    - **Validity:** H-0; M-6; L-2; I-0 **Pass**

Subgroup 3

- #2902 Contraceptive Care – Postpartum (HHS Office of Population Affairs)
    - **Reliability:** H-2; M-6; L-0; I-0 **Pass**
    - **Validity:** H-0; M-5; L-3; I-0 **Pass**
- #2903 Contraceptive Care – Most & Moderately Effective Methods (HHS Office of Population Affairs)
    - **Reliability:** H-5; M-3; L-0; I-0 **Pass**
    - **Validity:** H-1; M-5; L-2; I-0 **Pass**
- #2904 Contraceptive Care – Access to LARC (HHS Office of Population Affairs)
    - **Reliability:** H-3; M-5; L-0; I-0 **Pass**
    - **Validity:** H-0; M-7; L-1; I-0 **Pass**
- #3501e Hospital Harm – Opioid-Related Adverse Events (IMPAQ International/CMS)
    - **Reliability:** H-2; M-5; L-0; I-1 **Pass**
    - **Validity:** H-1; M-6; L-1; I-0 **Pass**

Measures for Discussion (Detailed)

Subgroup 1

#2880 Excess Days in Acute Care (EDAC) After Hospitalization for Heart Failure (HF)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The measure assesses days spent in acute care within 30 days of discharge from an inpatient hospitalization for HF to provide a patient-centered assessment of the post-discharge period. This measure is intended to capture the quality of care transitions provided to discharged patients who had a HF hospitalization by collectively measuring a set of adverse acute care outcomes that can occur post-discharge: emergency department (ED) visits, observation stays, and unplanned readmissions at any time during the 30 days post-discharge. In order to aggregate all three events, we measure each in terms of days. The Centers for Medicare & Medicaid Services (CMS) annually reports the measure for patients who are 65 years or older, are enrolled in Medicare Fee-For-Service (FFS), and are hospitalized in non-federal short-term acute care hospitals.
- **Type of measure:** Outcome
- **Data source:** Claims, Other: Census Data/American Community Survey, VHA Administrative Data, Medicare Enrollment Data (including the Master Beneficiary Summary File)
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 37 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-0; M-8; L-1; I-0 → Measure passes with MODERATE rating.
    - Reliability testing was conducted at the measure score level:
        - The developer used split-sample reliability without replacement (creating two non-overlapping samples) to assess reliability comparing intraclass correlation coefficients (ICCs) for hospitals with varying numbers of admissions.
        - The split sample from ICCs range from 0.465-0.698, with roughly two-thirds of the hospitals (i.e., those with ≤50 admissions) having ICC values less than or equal to 0.60. For hospitals with greater than or equal to 25 admissions, the ICC is 0.53. Hospitals with greater than or equal to 300 admissions had an ICC value of approximately 0.70.
- **Ratings for validity:** H-1; M-3; L-2; I-3 → Consensus not reached
    - Validity testing was conducted at the measure score level:
        - Face validity was assessed using survey-based information provided by the 16-member Technical Expert Panel (TEP).
            - Greater than 80 percent of experts moderately or strongly agreed with the validity of the measure.
        - Construct validity was assessed as the relationships between the HF EDAC measure score and three other measures: (1) the Hospital Star Rating readmission group scores, (2) the overall hospital Star Rating scores, and (3) the HF readmission measure.
        - The developer posited a negative relationship between the HF EDAC scores, star-rating readmission score group, and star-rating summary scores. They also

hypothesized a positive relationship between the HF EDAC scores and the HF RSRR score:

- Correlation Between HF EDAC Scores and Star Rating Readmission Group Scores -0.418 (p<.0001)
- Correlation Between HF EDAC Scores and Overall Hospital Star Rating Scores -0.371 (p<.0001)
- Correlation Between HF EDAC Scores and HF Readmission Measure Scores 0.574 (p<.0001)

- Regarding risk adjustment, the developer found a c-statistic of 0.59 and an R2 value of 0.027. Two social risk factors were tested and found to be statistically significant but do not appear to meaningfully affect hospital performance estimates; therefore, they were not included.
- Some SMP members raised concerns with the c-statistic, noting that it was low and that there was inadequate testing of the impact of leaving social risk factors out of the final model. SMP members raised additional concerns regarding the inadequate accounting for clustering within patients within the model.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
  - The SMP should discuss the appropriateness of the risk adjustment model and re-vote on validity.

#2881 Excess Days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** Measure score: The measure is a risk standardized score at the hospital level for days spent in acute care for patients with an AMI.

  Measure focus and time frame: This measure estimates days spent in acute care (i.e., time spent in ED, unplanned readmission and observation stays) within 30 days of discharge from an inpatient hospitalization for acute myocardial infarction (AMI). This measure is intended to capture the quality of care transitions provided to discharged patients hospitalized with AMI by collectively measuring a set of adverse acute care outcomes that can occur post-discharge: 1) emergency department (ED) visits, 2) observation stays, and 3) unplanned readmissions at any time during the 30 days post-discharge. Readmissions are classified as planned and unplanned by applying the planned readmission algorithm (PRA). Days spent in each care setting are aggregated for the 30 days post-discharge with a minimum of half-day increments (i.e., an ED visit lasting 2 hours would be counted as 0.5 days).

  Target population: 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: Census Data/American Community Survey, VHA Administrative Data, Medicare Enrollment Data (including the Master Beneficiary Summary File)
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 31 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-0; M-4; L-5; I-0 → Consensus not reached
    - Reliability testing was conducted at the measure score level:
        - The developer used split-sample reliability without replacement (creating two non-overlapping samples) to assess reliability comparing ICCs for hospitals with varying numbers of admissions.
            - The split sample reliabilities ranged from 0.23 – 0.63, with roughly three-fourths of the hospitals (i.e., those with ≤50 admissions) having ICC values less than or equal to 40. The ICC was 0.38 for hospitals with greater than or equal to 25 admissions and 0.63 for hospitals with greater than or equal to 300 admissions.
            - Some SMP members raised concerns with the low ICC scores, noting that the ICC values are generally less than or equal to 40 for the majority of hospitals, including 0.38 for greater than or equal to 25 admissions, indicating low/moderate agreement between the split samples.
- **Ratings for validity:** H-0; M-3; L-3; I-3 → Measure did not pass
    - Validity testing was conducted at the measure score level:
        - Face validity was assessed using survey-based information provided by the 16-member TEP.
            - Greater than 80 percent of experts moderately or strongly agreed with the validity of the measure.
        - Construct validity was assessed as the relationships between the acute myocardial infarction (AMI) EDAC measure score and three other measures: (1) the Hospital Star Rating readmission group scores, (2) the overall hospital Star Rating scores, and (3) the AMI readmission measure.
        - The developer posited a negative relationship between the MI EDAC scores, star-rating readmission score group, and star-rating summary scores. They also hypothesized a positive relationship between the MI EDAC scores and the MI RSRR scores:
            - Correlation Between AMI EDAC Scores and Star Rating Readmission Group Scores -0.260 (p<.0001)
            - Correlation With Overall Hospital Star Rating summary score -0.228 (p<.0001)
            - Correlation With AMI Readmission Measure: 0.425 (p<.0001)
        - Regarding risk adjustment, the developer found a c-statistic of 0.60 and an R2 value of 0.061. Two social risk factors were tested and found to be statistically significant but do not appear to meaningfully affect hospital performance estimates; therefore, they were not included.
        - Some SMP members raised concerns with the c-statistic, noting that it was low.

- Other SMP members stated that the choice of variables for the construct validity was inappropriate.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
  - The SMP should review and discuss the issues raised, including the developer's responses related to reliability, and re-vote on reliability.
  - The SMP should review and discuss the issues raised, including the developer's responses related to the appropriateness of the risk adjustment model, and re-vote on validity.

#2882 Excess Days in Acute Care (EDAC) After Hospitalization for Pneumonia

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure assesses days spent in acute care within 30 days of discharge from an inpatient hospitalization for pneumonia, including aspiration pneumonia or for sepsis (not severe sepsis) with a secondary discharge diagnosis of pneumonia coded in the claim as present on admission (POA) and no secondary diagnosis of severe sepsis coded as POA. This measure is intended to capture the quality of care transitions provided to discharge patients hospitalized for an eligible pneumonia condition by collectively measuring a set of adverse acute care outcomes that can occur post-discharge: emergency department (ED) visits, observation stays, and unplanned readmissions at any time during the 30 days post-discharge. In order to aggregate all three events, we measure each in terms of days. The Centers for Medicare & Medicaid Services (CMS) annually reports the measure for patients who are 65 years or older, are enrolled in Medicare fee-for-service (FFS), and are hospitalized in non-federal short-term acute care hospitals.
- **Type of measure:** Outcome
- **Data source:** Claims, Enrollment data
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 41 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-1; M-8; L-0; I-0 → Measure passes with a MODERATE rating
  - Reliability testing conducted at the measure score level:
    - The developer used split-sample reliability without replacement (creating two non-overlapping samples) to assess reliability comparing intraclass correlation coefficients (ICCs) for hospitals with varying numbers of admissions.
    - The split sample reliabilities ranged from 0.57 to 0.70. For hospitals with greater than or equal to 25 admissions, the ICC equals 0.576, for greater than or equal to 50 admissions, the ICC equals 0.628, and for greater than or equal to 300 admissions, the ICC equals 0.709.
    - No major concerns were raised by the SMP.
- **Ratings for validity:**H-0; M-3; L-3; I3 → Measure does not pass
  - Validity testing conducted at the measure score level:
    - Face validity was assessed using survey-based information provided by the 16-member Technical Expert Panel.

        - Greater than 80 percent of experts moderately or strongly agreed with the validity of the measure.
    - Construct validity was assessed as the relationships between the pneumonia (PN) EDAC measure score and the risk-standardized readmission rate (RSRR) group scores, the overall hospital rating scores, and the PN readmission measure.
    - The developer posited a negative relationship between the PN EDAC scores, star-rating readmission score group, and star-rating summary scores. They also hypothesized a positive relationship between the pneumonia EDAC scores and the pneumonia RSRR scores.
        - Correlation with Hospital Star Rating readmission group score: -0.416 (p<.0001)
        - Correlation with Overall Hospital Star Rating summary score: -0.398 (p<.0001)
        - Correlation with PN Readmission Measure: 0.625 (p<.0001)
    - Regarding risk adjustment, the developer found a c-statistic of 0.62 and an R2 value of 0.038. Two social risk factors were tested and found to be statistically significant but do not appear to meaningfully affect hospital performance estimates; therefore, they were not included.
    - Some SMP members raised concerns with the c-statistic, noting that it was low. Additionally, one SMP member commented that the choice of variables for testing construct validity were inappropriate.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
    - The SMP should review and discuss the issues raised, including the developer's responses, related to the appropriateness of the risk adjustment model and revote on validity.

#3188 30-Day Unplanned Readmissions for Cancer Patients

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The 30-Day Unplanned Readmissions for Cancer Patients measure is a cancer-specific measure. It provides the rate at which all adult cancer patients covered as Fee-for-Service Medicare beneficiaries have an unplanned readmission within 30 days of discharge from an acute care hospital. The unplanned readmission is defined as a subsequent inpatient admission to a short-term acute care hospital, which occurs within 30 days of the discharge date of an eligible index admission and has an admission type of "emergency" or "urgent."
- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 11 risk factors (with 15 values)
- **Sampling allowed**: None
- **Ratings for reliability:** H-0; M-7; L-2; I-0 → Measure passes with a MODERATE rating

  - Reliability testing conducted at the performance measure score level:
    - The developer conducted 250 split-sample iterations to assess reliability and used the 50 admissions threshed.
    - The median ICC score is 0.528 (mean: 0.527, 95% CI: 0.489, 0.558) in the risk adjusted split-sample tests.
    - In general, the SMP members agreed that the split sample approach is adequate, and the results are acceptable.
- **Ratings for validity:** H-0; M-3; L-4; I-2 → Measure does not pass
  - Validity testing conducted at both the critical data element and measure score levels:
    - The developer created a crosswalk of ICD-9 to ICD-10 for the codes used to define the denominator, denominator exclusions, numerator exclusions, and certain risk adjustment variables.
    - The correlation between *30-Day Unplanned Readmission for Cancer Patients* and CMS' *Hospital-Wide All-Cause Readmission Measure (HWR)* (NQF #1789) was 0.255 (95% CI 0.218, 0.291) ($p<0.001$), for the 2,412 hospitals, for which data on both measures were available.
    - Several SMP members raised concerns about the endogeneity nature of the correlation analysis; that is, the denominator for the HWR measure includes patients with cancer, and the same readmissions are in the numerators of both measures. The validity analysis only shows that the measure is correlated with itself (i.e., readmission) rather than correlated with quality.
    - There are also several concerns about the risk adjustment approach:
      - Two of these risk factors (e.g., Length of Stay (LOS) >3 days and ICU stay) are not present at the start of care, and there is a well-studied pathway connecting shorter stays with higher readmission risk (although the developers find the opposite association)
      - The overall c-statistic (c=0.711) is acceptable, but calibration in the highest risk group seems poor.
      - The rationale for excluding race seems to contradict the rationale for including dual eligibility.
      - Key data element testing should include discharge status field and risk adjustment variables.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
  - Discuss the various issues raised by the SMP members regarding validity and revote on validity.

#3612 Risk-Standardized Acute Cardiovascular-Related Hospital Admission Rates for Patients With Heart Failure Under the Merit-Based Incentive Payment System

MEASURE HIGHLIGHTS

- New Measure
- **Description:** Risk-standardized rate of acute, unplanned cardiovascular-related hospital admissions among Medicare Fee-for-Service (FFS) patients aged 65 years and older with heart failure (HF) or cardiomyopathy.

- **Type of measure:** Outcome
- **Data source:** Claims, Other: Medicare Fee-for-Service (FFS) administrative claims data, Medicare Enrollment Database (EDB), 2014 United States Department of Agriculture Economic Research Service, 2016 Area Health Resources File (AHRF), 2013-2017 American Community Survey (ACS), MIPS eligible provider files
- **Level of analysis:** Clinician: Group/Practice, Clinician: Individual
- **Risk-adjusted:** Statistical risk model with 30 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-0; M-5; L-4; I-0 → Consensus not reached
    - Reliability testing was conducted at the performance score level:
        - The developer performed a signal-to-noise analysis.
        - The minimum HF patient sample size for TINs needed to achieve **minimum** reliability scores of 0.4 and 0.5 (i.e., within the range of adequate reliability), to be between 21 and 32.
    - The SMP members agreed that approach is appropriate, but they raised several concerns listed below:
        - The reliability tests are not conducted and presented for clinical groups and individual clinicians separately.
        - The unit of analysis is not clear throughout the result section.
        - It is unclear whether the ICC should be interpreted as squared correlations between estimated and true values.
        - Reliabilities are much higher (median = 0.60) among providers with at least 21 eligible cases. However, only 24 percent of providers had this many cases, which again sounds a bit low.
        - It is also not clear whether the methods used to estimate reliability have taken provider/patient ratios into account (i.e., whether higher volume practices have a large number of clinicians each seeing a small number of patients vs. individual providers with a high volume of patients).
- **Ratings for validity:** H-1; M-4; L-2; I-2 → Consensus not reached
    - Validity testing was conducted at the performance measure score level:
        - Face validity is demonstrated through assessment of measure face validity by external groups, a TEP), and use of established measure development guidelines.
        - Of the 13 Clinician Committee members who responded to the survey, 11 of them, or 85 percent, strongly, moderately, or somewhat agreed that the MIPS HF measure can be used to distinguish good from poor quality of care.
        - The SMP members raised some concerns about the clarity of measure specifications, including attribution, exclusions (e.g., patients in hospice, patients with no Outpatient Evaluation and Management (E&M) visits, CKD-4), and whether HF is the primary diagnosis.
        - The SMP members, in general, thought the face validity is established adequately, although the TEP's feedback did not appear especially persuasive.
        - In general, the SMP members thought the risk adjustment model is adequate, although they noticed that indicators of HF severity are not included, and the model did not appear to account for the repeated measures' impact (i.e., single

patient with multiple admissions vs. multiple patients with single admission). SMP members also had questions about why race and dual eligibility are not included, as they both can affect the outcome.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
  - The SMP will discuss several issues regarding reliability (e.g., unit of analysis, result, etc.) and re-vote on reliability.
  - The SMP will discuss several issues regarding validity and re-vote on validity.

#3615 Unsafe Opioid Prescriptions at the Prescriber Group Level (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** Percentage of all dialysis patients attributable to an opioid prescriber's group practice who had an opioid prescription written during the year that met one or more of the following criteria: duration >90 days, Morphine Milligram Equivalents (MME) >50, or overlapping prescription with a benzodiazepine.
- **Type of measure:** Process
- **Data source:** Claims, Registry data, Other: IDR Medicare Provider
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with 178 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-6; M-1; L-1 I-1; → Measure passes with HIGH rating
  - The developer used CROWNWeb, Medicare Claims, the CMS Medical Evidence form 2728, Medicare Part D Claims as data sources to test the measure. The analysis included 103,157 physicians in 5,123 groups (range: 1-2,328 clinicians) with an average of 40 patients per group (range: 11-2,411).
  - Physician groups must have more than 10 eligible patients to be included in the measure or the analysis.
  - Reliability testing was conducted at the score level:
    - Inter-unit reliability (IUR) calculated at the group level: 0.86
    - Profile IUR (PIUR): 0.98
  - SMP Comments: Specifications
    - "Minor: … it wasn't clear to me whether prevalent comorbidities in the risk adjustment model were ascertained from claims submitted during the measurement period or some prior period."
    - "The developers did not include value sets for opioids or benzodiazepines. I am also unclear about the duration of the reporting period, as it is not clearly stated anywhere (presumably one year). Finally, it is unclear what happens when multiple prescriptions from multiple prescribers collectively contribute >90 days; do all prescribers get blamed, even the first in the sequence of prescriptions? The one-month minimum duration of Part D enrollment is not consistent with the duration component of the numerator."

        - "Two issues regarding measure specifications: [1] The MIF (in S.7) notes the denominator includes "dialysis patient[s]". However, this is not defined here nor in the XL data dictionary file. [2] The MIF (in S.8) notes an exclusion of "Patients who have a hospice claim". In S.9 we only receive a high level definition of hospice, but it's definitely not clear how this is exclusion is defined based on this response. Example: Unstated as to what "CMS file" is referenced. Unstated what fields in said form & what responses / codes are employed to operationally meet the hospice definition so as to exclude a given case."
        - "I'm struggling with the distinction between prescriber group and provider group."
    - SMP Comments: Reliability testing
        - Most panelists suggested that the testing approach was appropriate and within acceptable limits.
        - "IUR and PIUR were used to estimate provider level reliability, and outlier effects for groups with at least 11 patients. The split sample analysis as described states that patients were divided into two equal groups within practice, and that the process was repeated 100 times. For small groups that process would yield uninterpretable results. The average number of patients per group was 40 and the average number of physicians per group was 20. It is not clear how the variable number of physicians per group was handled in this process."
- **Ratings for validity:** H-2; M-4: L-1; I-2 → Measure passes with MODERATE rating
    - Validity testing was conducted at the score level:
        - The developer conducted a concordance analysis of the relationship between measure scores, hospitalization, and mortality.
        - Hospitalization rate at the practitioner group level is 1.49, 1.46 and 1.41 for T1, T2, and T3 respectively (trend test $p<0.001$), while the average number of hospital days per year and patient at the practitioner group level is 6.1, 5.1 and 4.1 respectively (trend test $p<0.001$).
        - The practitioner group level average mortality rate is 0.19, 0.20, and 0.18 per patient-year for T1, T2, and T3 groups, respectively.
    - SMP Comments: Validity testing
        - "The main validity questions pertain to the choice of numerator criteria and case mix adjustment. I would like to know that there is ample evidence/consensus that numerator occurrences necessarily reflect a lapse in care quality (as opposed to rare cases where such prescribing is actually appropriate) and that the measure adequately adjusts for potential differences in pain severity."
        - "Assessing the correlation between 3615 and hospitalization and mortality is an appropriate validity test given the provided discussion of the literature. However, there is not specific correlation test specified. It appears the relationships are stated with descriptive statistics."
    - SMP Comments: Risk adjustment
        - Most SMP members agreed with developer's approach.

- "The developers discuss the rationale for various modeling choices, but they didn't say much about the decision to adjust in the first place. I think it makes sense, but I could also imagine arguments against adjustment. The model adjusts for age, sex, BMI, duration of ESRD, nursing home status in previous year, diabetes as primary cause of ESRD, comorbidities at ESRD incidence, and prevalent comorbidities. It wasn't clear to me whether prevalent comorbidities are assessed based on claims submitted during the measurement period or during a time interval preceding the measurement period. If the former, then strictly speaking model is violating the principle that adjustment variables should be present before the start of care. This concern is probably more theoretical than practical. The model appears to be well calibrated overall. I was curious to know if the developers assessed calibration within subgroups such as patients with multiple comorbidities. Assessing calibration by race could shed light on the potential consequences of not adjusting for race despite the apparent strong odds ratios for race categories. The calibration plot Figure 4 compares observed versus expected based on absolute numbers of patients instead of proportions (probabilities). I was curious to know if calibration looked equally good when plotting proportions."
- "Risk-adjustment approach is very poorly justified. Process measures are rarely risk-adjusted, because the presumption is that "safe and effective" care is linked to a specific denominator-eligible population (after exclusions and stratification, as appropriate). A complex risk-adjustment model of this type belies the concept of "unsafe opioid prescriptions." Implicit in their risk-adjustment approach is the concept that "unsafe prescribing" must be safe and effective for some subsets of the eligible population, or else why would we account for these patient characteristics in a process measure? Their models include age, sex, BMI, time on ESRD, nursing home residence, cause of ESRD, and hundreds of comorbid conditions. Some of these covariates, like age, make clinical sense (i.e., it is clearly safer to prescribe medium-dose opioids to younger patients than to older patients). But how does the duration of ESRD or long-term NH residence, for example, "justify" so-called "unsafe opioid prescriptions"? Even worse, the list of covariates includes features that don't make clinical sense in this context, such as "renal failure," "chronic kidney disease," and "drug dependence."

ITEMS TO BE DISCUSSED

- **Action items:**
  - To what extent is the validity analysis confounded by unmeasured case mix, considering that dialysis physicians with sicker patients (e.g., such as those with comorbid cancer) have higher mortality rates, hospitalization rates, and opioid use?

#3616 Unsafe Opioid Prescriptions at the Dialysis Practitioner Group Level (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- New Measure

- **Description:** Percentage of all dialysis patients attributable to a dialysis provider's group practice who had an opioid prescription written during the year that met one or more of the following criteria: duration >90 days, Morphine Milligram Equivalents (MME) >50, or overlapping prescription with a benzodiazepine.
- **Type of measure:** Process
- **Data source:** Claims, Registry data, Other: IDR Medicare Provider
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with 178 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-1; M-6; L-1;I-1; → Measure passes with MODERATE rating
    - The developer used CROWNWeb, Medicare Claims, the CMS Medical Evidence form #2728, and Medicare Part D Claims as data sources to test the measure. A total of 6784 physicians in 3323 groups (range: 1-51 clinicians) with an average of 46 patients per group (range: 11-1022) were included in the analysis.
    - Physician groups must have more than 10 eligible patients to be included in the measure or the analysis.
    - Reliability testing was conducted at the score level:
        - IUR calculated at the group level: 0.60
        - PIUR: 0.81
    - SMP Comments: Specifications
        - "Minor: … it wasn't clear to me whether prevalent comorbidities in the risk adjustment model were ascertained from claims submitted during the measurement period or some prior period."
        - "Two issues regarding measure specifications: [1] The MIF (in S.7) notes the denominator includes "dialysis patient[s]". However, this is not defined here nor in the XL data dictionary file. [2] The MIF (in S.8) notes an exclusion of "Patients who have a hospice claim". In S.9 we only receive a high level definition of hospice, but it's definitely not clear how this is exclusion is defined based on this response. Example: Unstated as to what "CMS file" is referenced. Unstated what fields in said form & what responses / codes are employed to operationally meet the hospice definition so as to exclude a given case."
        - "The developers did not include value sets for opioids or benzodiazepines. I am also unclear about the duration of the reporting period, as it is not clearly stated anywhere (presumably one year). The one-month minimum duration of Part D enrollment is not consistent with the duration component of the numerator. Finally, the conceptual rationale for attributing all opioid prescriptions for dialysis patients to the dialysis practice group is not clearly articulated (although it would belong in the material reviewed by the Standing Committee, not the SMP)."
        - "Not clear between #3615 and #3615 how the group definitions are really different. There is a difference, but it is not clear and obvious."
    - SMP Comments: Reliability testing
        - Most panelists suggested that the testing approach was appropriate and within acceptable limits.

        - "The variation between providers within provider group does not appear to have been handled by the methods reported (i.e., the error term appears not to include between providers across patients within practice)."
- **Ratings for validity:** H-1; M-5; L-1; I-2; → Measure passes with MODERATE rating
    - Validity testing was conducted at the score level:
        - The developer conducted a concordance analysis of the relationship between measure scores, hospitalization, and mortality.
        - The hospitalization rate at the dialysis provider group level is 1.55, 1.48, and 1.47 for tercile 1 (T1), T2, and T3, respectively (trend test p<0.001), while the average number of hospital days per year and patient at the dialysis provider group level is 8.3, 7.5, and 7.7, respectively (trend test p<0.001).
        - The dialysis provider group level average mortality rate is 0.26, 0.29, and 0.33 per patient-year for T1, T2, and T3 groups, respectively.
    - SMP Comments: Validity testing
        - "Assessing the correlation between 3615 and hospitalization and mortality is an appropriate validity test given the provided discussion of the literature. However, there is not specific correlation test specified. It appears the relationships are stated with descriptive statistics."
        - "Developers estimated differences across tertiles of performance in patient hospitalization rates and mortality rates. This is a promising approach, but the observed associations could be explained entirely by case mix/severity."
    - SMP Comments: Risk adjustment
        - Most SMP members agreed with developer's approach.
        - "Careful analysis of both clinical and social risk factors. Clinical factors retained and social factors eliminated, even though both have similar levels of predictive power. This could be a "fail" issue, but many developers are doing this in spite of 2014 Expert Panel recommendations to treat clinical and social variables the same."
        - "Risk-adjustment approach is very poorly justified. Process measures are rarely risk-adjusted because the presumption is that "safe and effective" care is linked to a specific denominator-eligible population (after exclusions and stratification, as appropriate). A complex risk-adjustment model of this type belies the concept of "unsafe opioid prescriptions." Implicit in their risk-adjustment approach is the concept that "unsafe prescribing" must be safe and effective for some subsets of the eligible population, or else why would we account for these patient characteristics in a process measure? Their models include age, sex, BMI, time on ESRD, nursing home residence, cause of ESRD, and hundreds of comorbid conditions. Some of these covariates, like age, make clinical sense (i.e., it is clearly safer to prescribe medium-dose opioids to younger patients than to older patients). But how does the duration of ESRD or long-term NH residence, for example, "justify" so-called "unsafe opioid prescriptions"? Even worse, the list of covariates includes features that don't make clinical sense in this context, such as "renal failure," "chronic kidney disease," and "drug dependence.""

ITEMS TO BE DISCUSSED

- **Action items:**
    - To what extent is the validity analysis confounded by unmeasured case mix, considering that dialysis physicians with sicker patients (e.g., comorbid cancer) have higher mortality rates, hospitalization rates, and opioid use?

#3622 National Core Indicators for Intellectual and Developmental Disabilities (ID/DD) Home and Community-Based Services (HCBS) Measures (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** National Core Indicators for Intellectual and Developmental Disabilities Home- and Community-Based Services Measures ("NCI for ID/DD HCBS Measures" hereafter) originate from NCI(R) In-Person Survey (IPS), an annual multi-state cross-sectional survey of adult recipients of state developmental disabilities systems´ supports and services. First developed in 1997 by the National Association of State Directors of Developmental Disabilities Services (NASDDDS) in collaboration with Human Services Research Institute (HSRI), the main aims of NCI for ID/DD HCBS Measures were to evaluate person-reported outcomes and assess state developmental disabilities service systems performance in various domains and sub-domains accordingly. The unit of analysis is "the state", and the accountable entity is the state-level entity responsible for providing and managing developmental disabilities services. Currently, 46 states and the District of Columbia are members of the NCI program. To align with member states´ fiscal schedules, the annual survey cycle typically starts on July 1 and ends on Jun 30 of the following year.
  Gathering subjective information and data from people with ID/DD poses unique challenges due to potential intellectual and developmental limitations experienced by the population. As such, extensive work went into the processes of developing NCI IPS administration methods, survey methodology and measure design and revisions. The original development built on direct consultation with members of the target population and their advocates, as well as extensive literature review and testing.
  The NCI for ID/DD HCBS Measures consist of 14 measures in total, including:
  Five measures in the HCBS Domain: Person-Centered Planning (PCP) and Coordination
  #PCP-1 The proportion of people who express they want a job who have a related goal in their service plan (Community Job Goal)
  #PCP-2 The proportion of people who report their service plan includes things that are important to them (Person-Centered Goals)
  #PCP-3 The proportion of people who express they want to increase independence in functional skills (ADLs) who have a related goal in their service plan (ADL Goal)
  #PCP-4 The proportion of people who report they are supported to learn new things (Lifelong Learning)
  #PCP-5 The proportion of people who report satisfaction with the level of participation in community inclusion activities (Satisfaction with Community Inclusion Scale)
  Four measures in the HCBS Domain: Community Inclusion
  #CI-1 The proportion of people who reported that they do not feel lonely often (Social Connectedness)
  #CI-2 The proportion of people who reported that they have friends who are not staff or family members (Has Friends)

#CI-3 The proportion of people who report adequate transportation (Transportation Availability Scale)
#CI-4 The proportion of people who engage in activities outside the home (Community Inclusion Scale)
Four measures in the HCBS Domain: Choice and Control
#CC-1 The proportion of people who reported they chose or were aware they could request to change their staff (Chose Staff)
#CC-2 The proportion of people who reported they could change their case manager/service coordinator (Can Change Case Manager)
#CC-3 The proportion of people who live with others who report they can stay home if they choose when others in their house/home go somewhere (Can Stay Home When Others Leave)
#CC-4 The proportion of people who report making choices (independently or with help) in life decisions (Life Decisions Scale)
And one measure in the HCBS Domain: Human and Legal Rights
#HLR-1 The proportion of people who report that their personal space is respected in the home (Respect for Personal Space Scale)

- **Type of measure:** Outcome: PRO-PM
- **Data source:** Instrument-based data
- **Level of analysis:** Population: Regional and State
- **Risk-adjusted:** Statistical risk model with 13 risk factors and stratification by five (residence type) risk categories
- **Sampling allowed**: Each state is instructed to construct a sample frame of adults (18 and over) who are receiving at least one publicly funded ID/DD service in addition to case management. Based on this sample frame and the assumption of a middle response distribution (50 percent), each state is recommended to have a sample size that will support both (1) a 95 percent confidence level and (2) a 5 percent margin of error. States are allowed to design their own stratifying strategy (e.g., stratifying by Medicaid waiver funding types), as long as they provide the information needed for weighting.
- **Ratings for reliability:** H-3; M-3; L-2; I-1; → Measure passes with MODERATE rating
    - Reliability testing was conducted at the data element level:
        - The developer describes multiple data element analyses conducted, some from previous work conducted and others based on a relatively recent sample of In-Person Surveys (IPS) of the National Core Indicators (NCI). Sample includes 37 states and a total of 22,000 completed surveys.
        - IRR studies summarized by percent agreement and Kappa statistics. Studies conducted in 1997, 1998, 1999, 2008, and 2010 resulted in the following average agreements/kappa scores:
            - 1997: average agreement of 93 percent;
            - 1998: average agreement of 93 percent/average kappa score of 0.79;
            - 1999: average agreement of 92 percent;
            - 2008: average kappa score of 0.90;
            - 2010: 80 percent/average kappa score of 0.89.
        - Cognitive tests for data elements were conducted to ensure respondent understanding. The number of valid responses to the items ranged between eight and 10 with an overall average of nine over all 35 items.

- Principal components analysis (PCA) exploratory factor analysis: The 5 items comprising the Life Decisions Scale constituted a single factor that explained 45 percent of the variance with component loadings ranging from 0.521 to 0.759.
    - The three items comprising the Respect for Personal Space Scale constituted a single factor that explained 44 percent of the variance with component loadings ranging from 0.490 to 0.746.
    - The two items comprising the Transportation Availability Scale constituted a single factor that explained 71 percent of the variance with both components loading at 0.842.
    - The four items comprising the Community Inclusion Scale constituted a single factor that explained 44 percent of the variance with component loadings ranging from 0.583 to 0.720.
    - The five items comprising the Satisfaction with Community Inclusion Scale constituted a single factor that explained 46 percent of the variance with component loadings ranging from 0.583 to 0.763.
- Internal consistency of scales
    - Cronbach's alpha for multi-item scales
        - Life Decisions Scale: 0.686
        - Respect for Personal Space Scale: 0.349
        - Community Inclusion Scale: 0.687
        - Satisfaction with Community Inclusion Scale: 0.704
    - Spearman-Brown coefficient for two-item scales: The Transportation Availability Scale was 0.591.
    - Corrected Item: Total Correlation Coefficients ranged from 0.134-0.551 across the 14 items.

- Reliability testing was conducted at the score level:
    - The developer conducted ANOVA to assess between-state variance in relationship to within-state variance and assessed inter-unit reliability (IUR). ANOVA analysis found that between-state variation is significantly larger than within-state variation for each of the 14 measures ($p<0.001$). The IUR ranged between 0.753 and 0.984.
- SMP Comments: Specifications
    - SMP members suggested that the developer clarify the sampling methodology, noting that "that each state is recommended to have a sample size that will support both a 95% confidence level and a ±5% margin of error, but not what measures those are based on, minimum recommended sample size, sampling procedures, etc. Without more detail there is concern that sampling procedures could vary substantially by state."
    - SMP members questioned if this is a PRO-PM.
- SMP Comments: Reliability testing approach
    - SMP members primarily noted the reliability testing approach was appropriate with few concerns.
    - One SMP member noted "issues with the sections / domains with test results: [1] They do not always match up with the 4 domains of 3622. Example: The MIF states a section is titled "Choice and Control". However, on p. 10 of the testing

form, the 2 sets of results on top of the page do not include results for the "Choice…" section. [2] They differ from various test results reported. Example: Page 10 has test results for section titled "Respect for Personal Space". However, test results on p. 9 do not report test results for this section. [3] The test results bottom p. 9 that have a crosswalk between section name & measure ID, these section names do not carry through in subsequent test results."

- **Ratings for validity:** H-0; M-2; L-3; I-4; → Measure does not pass with LOW/INSUFFICIENT rating
    - Validity testing was conducted at the data element level:
        - The developer suggests that interviewers are asked to give formal feedback on interviews conducted to ensure individual interview validity.
        - The developer provides a list of seven references for studies investigating the relationships among NCI data elements and testing hypotheses about expected associations.
    - Validity testing was conducted at the score level:
        - The developer reports Pearson Product Moment Correlation Coefficients among the 37 states performance scores between the 14 IPS items.
        - The score range from 0.345-0.763 suggests a moderate to high correlation result.
    - SMP Comments: Validity testing approach
        - "The developers do not provide clear evidence regarding data element validity, so this evaluation is based on measure score validity. They apparently did not do formal testing of content validity, or else they declined to report the results. They describe assessments of "the individual's comprehension of the questions and consistency of responses," but they declined to provide results of these assessments. They cite to "multiple published studies (that) investigated relationships among NCI data elements," but they do not summarize any key findings from these studies. Their testing of performance score reliability at the state level is reasonable, but not optimal, because all of the constructs are estimated based on the same survey. Therefore, any validity issues that affect the entire survey in a consistent manner are likely to lead to exaggerated correlations."
        - "Principal components analysis was performed on individual multi-item scales and results presented appear adequate, however statistics appropriate model (e.g. CFI, RMSEA, etc.) were not provided, nor were the actual factor analysis. Further, more helpful to demonstrate convergent/discriminant validity evidence would have been a rotated factor analysis (e.g., varimax or obliques) of all multi-item scales together. The results of the Pearson product-moment correlation analysis are difficult to interpret since the theoretical relationship to the correlates chosen was not provided, whether only significant associations were returned/presented, etc. Also, Bonferroni correction for significance should have been done, particularly if a balanced correlation across all validation variables was performed and only significant ($p<0.5$) values retained."

- "Regarding testing of critical data elements: Given the measure is risk adjusted, I would consider the risk factors as critical data elements. In this regard, none of the risk factors were tested."

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
  - The SMP will review the developer's responses to validity issues and re-vote on validity.

Subgroup 2

#3614 Hospitalization After Release With Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This outcome measure tracks the rate of patients admitted to the hospital for a stroke within 30 days of being treated and released from the ED with either a non-specific, presumed benign symptom-only dizziness diagnosis or a specific inner ear/vestibular diagnosis (collectively referred to as "benign dizziness"). The measure accounts for the epidemiologic base rate of stroke in the population under study using a risk difference approach (observed [short-term rate] minus expected [long-term rate]).
- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Facility
- **Not risk-adjusted**
  - The rationale for no risk adjustment is that the developer uses a statistical risk approach (observed short-term stroke risk within 30 days minus expected). The short-term expected rate is estimated in the same patients using the 30-day rate of stroke admission during the period of 91-360 days post-discharge. This attempts to quantify the excess short-term risk of stroke (i.e., attributable risk) due to misdiagnosis. This is not risk-adjusted because, according to the developer, it captures all hospital and patient characteristics and social risk factors.
- **Sampling allowed**: None
- **Ratings for reliability:** H-0; M-5; L-1; I-;2 → Measure passes with MODERATE rating
  - Reliability testing was conducted at the measure score level:
    - A signal-to-noise analysis was conducted.
    - This sampling strategy of restricting to hospitals with over 250 cases does not match with how the measure is intended to be implemented.
    - The median reliability score for the entire 967-hospital sample was 0.590 with an interquartile range of 0.414-0.951. Reliability was described as only "okay". It was marginal/low in hospitals with 250-499 benign dizziness ED discharges (0.582). Results show moderate reliability for hospitals in hospitals that had over 250 dizziness discharges over the three-year time frame. The measure was not reliable below this level of cases. Some comments stated it would not be reliable unless over 500 cases.
    - There was concern regarding a large number of facilities with a reliability score of 1.0; it is unclear whether they had no negative events.

        - There were concerns that for the denominator definition, a patient may have multiple "index" visits over a three-year period and that this was not accounted for in the reliability testing.
        - There was a concern that only 967 out of 5,503 facilities were included, which would only include EDs with volumes of over 40,000 discharges per year, which would allow for the 250 threshold benign dizziness cases over the three-year period.
- **Ratings for validity:** H-2; M-2; L-3; I-1; → Consensus not reached
    - Validity testing was conducted at the data element level:
        - Data element validity was assessed for two reasons: (1) to test whether stroke diagnoses were valid and (2) to test whether claims were intended to be coded as "benign dizziness" by the clinician when they were coded as such.
        - For data element validity for stroke, the developers cited prior literature that used claims data to identify stroke discharges using chart abstraction as the standard. In this approach, there was a sensitivity for stroke of 86 percent, specificity of 95 percent, PPV of 90 percent, with a kappa agreement of 0.82. In a systematic review of 77 studies, the sensitivity for any cerebrovascular disease was greater than 82 percent in most studies and both specificity and NPV were greater than 95 percent.
        - For denominator reliability for benign dizziness diagnoses, they conducted two studies focused on code-level validity. First, when an ED patient has a "benign dizziness" discharge diagnosis, how often do charts suggest the ED provider INTENDED to code "benign dizziness"? This was conducted using two academic hospitals. PPV was calculated in a random sample of 64 charts in three cohorts (i.e., chief complaints of dizziness, oto-vestibular complaints, and other chief complaints). Second, they calculated an NPV specifically if another diagnosis was coded; how often did they intend to code something other than benign dizziness? They reviewed a random sub-sample of 67 charts for high-risk sub-group to estimate NPV. The PPV was 100 percent for coding benign dizziness. The NPV was nearly 100 percent. The audit of discharged status demonstrated 100 percent accuracy, even for the highest risk cases.
        - The observed rate of stroke in 30 days among cases was compared to an "expected rate" to calculate the measure, the latter being 91-360 days after the visit.
        - There were concerns that the expected rate was based on the assumption that the risk of stroke in 91-360 days is not associated with a misdiagnosis of benign dizziness. The SMP also raised concerns that this approach to calculate the "expected rate" fully accounts for risk factors of patients.
        - The lack of risk adjustment for social risk factors was only mentioned in the context of "risk-adjusting away" worse care for racial minorities, and no discussion of potential conceptual relationships.
        - Only a limited sample of hospitals (four hospitals within Johns Hopkins) were used for testing, which may not generalize to other hospitals.
        - Only a very small number of hospitals are extremely poor performers, eight out of 927, suggesting that this is a rare event.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
    - The SMP will review the developer's responses to validity issues and re-vote on validity.

Subgroup 3

#0500: Severe Sepsis and Septic Shock: Management Bundle (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure focuses on adults 18 years and older with a diagnosis of severe sepsis or septic shock. Consistent with Surviving Sepsis Campaign guidelines, it assesses measurement of lactate, obtaining blood cultures, administering broad spectrum antibiotics, fluid resuscitation, vasopressor administration, reassessment of volume status and tissue perfusion, and repeat lactate measurement. As reflected in the data elements and their definitions, the first three interventions should occur within three hours of presentation of severe sepsis, while the remaining interventions are expected to occur within six hours of presentation of septic shock.
- **Type of measure:** Composite
- **Data source:** Electronic Health Data, Paper Medical Records
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: Hospitals have the option to sample from their population or submit their entire population. Hospitals whose Initial Patient Population size is less than the minimum number of cases per quarter/month for the measure cannot sample.
  Hospitals that choose to sample have the option of sampling quarterly or sampling monthly. A hospital may choose to use a larger sample size than is required. Hospitals whose Initial Patient Population size is less than the minimum number of cases per quarter/month cannot sample. Hospitals that have five or fewer sepsis discharges for the entire measure set (both Medicare and non-Medicare combined) in a quarter are not required but are encouraged to submit sepsis patient level data to the CMS Clinical Warehouse.
  Hospitals selecting sample cases for the sepsis measure must ensure that the population and quarterly sample size meets the following conditions:
    - If average quarterly initial patient population size "N" is greater than or equal to 301, then the minimum required sample size is 60.
    - If average quarterly initial patient population size "N" is 151-300, then the minimum required sample size is 20 percent of the initial patient population size.
    - If average quarterly initial patient population size "N" is 30-150, then the minimum required sample size is 30.
    - If average quarterly initial patient population size "N" is six to 29, then there is no sampling; one hundred percent of the initial patient population is required.
    - If there are zero to five cases, then submission of patient level data is encouraged but not required. If submission occurs, one to five cases of the Initial Patient Population may be submitted.

Hospitals selecting sample cases for the sepsis measure must ensure that the population and monthly sample size meets the following conditions:

  - If average quarterly initial patient population size "N" is greater than or equal to 101, then the minimum required sample size is 20.
  - If average quarterly initial patient population size "N" is 51-100, then the minimum required sample size is 20 percent of the initial patient population size.
  - If average quarterly initial patient population size "N" is 10-50, then the minimum required sample size is 10.
  - If average quarterly initial patient population size "N" is less than 10, then there is no sampling; one hundred percent of the initial patient population is required.

- **Ratings for reliability:** H-5; M-1; L-0; I-2; → Measure passes with a HIGH rating
  - Reliability testing was conducted at the measure score level:
    - Signal-to-noise analyses with beta-binomial were conducted.
    - Two levels of measure score testing were conducted, one for all facilities, regardless of N, and the second level was only for facilities with a 10-case minimum. The latter represented 86 percent of the total.
    - There were concerns that data element reliability was not conducted for the individual elements of the measure.
    - The reliability score was 0.92 with a confidence interval of 0.41-1.00 for Q4 2015, an interval of 0.93 with a confidence interval of 0.47 - 1.00 for Q1 2016, and 0.93 with a confidence interval of 0.42 - 1.00 for Q2 2016. A change between 2015 to 2016 was noted, which then remained stable. For all facilities with 10 or more cases, the results 0.63-0.99 for Q42015, 0.64-0.99 for Q12016, and 0.65-0.99 for Q22016. It is noted that the range of the confidence interval tightened for the facilities with 10 or more cases.
- **Ratings for validity:** H-3; M-2; L-1; I-2; → Measure passes with a HIGH rating
  - Validity testing was conducted at the measure score and data element level:
    - For data element validity, abstracted values were compared to a gold standard for trained abstractors. There was moderate or high agreement for the majority of the data elements.
    - For measure score validity, the measure was compared to mortality rates using chi-squared testing, but there were concerns of aggregation bias. There was a strong inverse relationship between measure performance and mortality, which is to be expected.
    - Low agreement was noted for several important time variables.
    - The SMP expressed mixed sentiments regarding whether validity testing was appropriate.
    - There were concerns that only two quarters of data were used in light of multiple changes made to the measure outside of the Q3-4 2018 data used for the analysis.
    - There were concerns regarding the lack of justification for not adjusting for social determinants of health (SDOH) or race.
- **Ratings for composite**: H-2; M-3; L-0; I-1; → Measure passes with a MODERATE rating

- Each element of the all-or-none measure is informed by the literature and aligns with the Surviving Sepsis Campaign. Each element is part of a sequence of care, making an "all or none" measure more meaningful than just assessing compliance with individual elements.
- The components have been selected from clinical practice guidelines but not empirically analyzed for parsimony or contributions of each.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
  - The SMP could consider discussing the following issue identified by a SMP member. Both reliability and validity analyses were conducted only using 2018 Q3-Q4 data when no changes were made to the measure specifications. The measure stewards did not address the fact that significant updates were made to this measure between initial endorsement and this submission. Given that this measure is used in a CMS payment program with significant financial implications, the impact of changes made on performance score was not assessed.

#0674: Percent of Residents Experiencing One or More Falls With Major Injury (Long Stay) (Pulled by NQF Staff)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure reports the percentage of long-stay residents in a nursing home who have experienced one or more falls resulting in major injury (defined as bone fractures, joint dislocations, closed head injuries with altered consciousness, or subdural hematoma) reported in the look-back period no more than 275 days prior to the target assessment. The long stay nursing home population is defined as residents who have received 101 or more cumulative days of nursing home care by the end of the target assessment period. This measure is based on data obtained through the Minimum Data Set (MDS) 3.0 OBRA, PPS, and/or discharge assessments during the selected quarter(s).
- **Type of measure:** Outcome
- **Data source:** Assessment Data
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: None
- **Ratings for reliability:** H-0; M-6; L-2; I-0; → Measure passes with a MODERATE rating
  - Reliability testing was conducted at the measure score and data element level:
    - Data element reliability was assessed using inter-rater reliability of the gold-standard nurse to gold-standard nurse & gold-standard nurse to facility nurse and calculated kappas.
    - The results showed that the inter-rater reliability was 0.967 for gold-standard to gold-standard and was 0.945 for facility nurse to gold standard. This suggests substantial agreement.
    - Measure score reliability was assessed using both signal-to-noise & split-half reliability.

        - The signal-to-noise reliability was only 0.45, which was likely due to the nature of the outcome being a rare event and sample size issues.
        - The split-half correlation for this measure was positive (r = 0.18, ρ = 0.18, p < .01), providing limited evidence of internal reliability.
        - There was concern regarding the data being from 2008.
- **Ratings for validity:** H-1; M-6; L-1; I-0; → Measure passes with a MODERATE rating
    - Validity testing was conducted at the measure score level:
        - Data element validity was conducted with inter-rater reliability testing as described above in the reliability section.
        - The developer correlated the measure with a few other measures of MDS quality measures and claims-based measures.
        - The empirical validity testing comparing the correlations to other measures showed weak correlations.
        - There was a concern that 25 percent of facilities are jumping three or more deciles in performance over a short interval, which may be an issue with poor reliability.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
    - Is the reliability sufficient to warrant the moderate rating, given the marginal signal-to-noise ratings and the concerns regarding the stability analysis?

#0679 Percent of High-Risk Residents With Pressure Ulcers (Long Stay) (Pulled by NQF Staff)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure reports the percentage of long-stay, high-risk, residents in a nursing home who have Stage II-IV or unstageable pressure ulcers on a selected target assessment in the target quarter. The long stay nursing home population is defined as residents who have received 101 or more cumulative days of nursing home care by the end of the target assessment period. A nursing home resident is defined as high-risk for pressure ulcer if they meet one or more of the following three criteria: 1. Impaired bed mobility or transfer, 2. Comatose, 3. Malnourished or at risk of malnutrition. This measure is based on data obtained through the Minimum Data Set (MDS) 3.0 OBRA, PPS, and/or discharge assessments during the selected quarter(s).
- **Type of measure:** Outcome
- **Data source:** Assessment Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Other. Sample restriction - this measure is restricted to residents who are at high risk for pressure ulcers. Residents are identified as high risk if they meet any of the following three criteria: (1) impaired in bed mobility or transfer, (2) comatose, or (3) active diagnosis of malnutrition [protein or calorie] identified, or the resident is at risk for malnutrition. (See denominator details for more information). This measure was originally developed as one of a pair of stratified pressure ulcer measures – one low-risk and one high-risk. The low-risk measure is no longer reported or maintained.

- **Sampling allowed:** None
- **Ratings for reliability:** H-0; M-6; L-2; I-0; → Measure passes with a MODERATE rating
    - Reliability testing was conducted at the measure score and data element level:
        - Data element reliability was conducted between the gold-standard nurse abstractor and the facility nurse abstractor, calculating a kappa score. The Kappa value was 0.92 for gold-standard v. gold-standard and 0.97 for facility nurse v. gold-standard.
        - Measure score reliability testing was done with signal-to-noise and split half reliability testing.
        - The split-half correlation for this measure was positive, and the relationship was moderate ($r = 0.33$, $\rho = 0.30$, $p < .01$), suggesting there is modest evidence of internal reliability.
        - The average signal-to-noise reliability score was 0.50. This suggests that the measure is moderately reliable in separating facility characteristics from variability within facility.
        - This testing was conducted in 2008.
        - Moderate/low ratings were given because of the weak/moderate reliability testing scores.
- **Ratings for validity:** H-2; M-4; L-2; I-0; → Measure passes a MODERATE rating
    - Validity testing was conducted at the measure score and data element level:
        - Data element reliability testing was used as validity testing for this measure and results are described above.
        - Measure score validity was conducted by correlating the measure score with other measures of quality across states. Spearman correlations ranged from -0.207 to +0.203 for the measure score with the other measures of quality mentioned above. Approximately 5.84 percent of the variation was between-state. Average interquartile range of state-level scores was 6.4 percentage points. Of interest was the note that 24.6 percent of facilities did not change deciles, over 25.7 percent changed one decile, 19.4 percent changed two deciles, and 30.4 percent changed three or more deciles. This is attributed to low-frequency events and the impact on one event on the decile assignment.
        - The developer argued not to adjust for some potential risk factors, such as age and race, to avoid unintended consequences.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
    - Is the reliability sufficient to warrant the moderate rating?

#3621 Composite Weighted Average for 3 CT Exam Types: Overall Percent of CT Exams for Which Dose Length Product Is at or Below the Size-Specific Diagnostic Reference Level (for CT Abdomen-Pelvis With Contrast/Single Phase Scan, CT Chest Without Contrast/Single Phase Scan and CT Head/Brain Without Contrast/Single Phase Scan)

MEASURE HIGHLIGHTS

- New Measure

- **Description:** Weighted average of 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan)
- **Type of measure:** Composite; process
- **Data source:** Registry Data
- **Level of analysis:** Clinician: Group/Practice, Facility
- **Not risk-adjusted**
- **Sampling allowed:** None
- **Ratings for reliability:** 5 high, 2 moderate, 0 low, and 1 insufficient → Measure passes with a HIGH rating.
    - Reliability testing was conducted at the measure score level:
        - Signal-to-noise analysis with beta-binomial was conducted to measure the confidence to differentiating between radiology groups.
        - The reliability score was above .997 for all types of CT's and the composite weighted average.
        - The developer indicated that the measure is specified for a facility but did not provide any facility-level reliability testing.
- **Ratings for validity:** H-0; M-4; L-0 I-4; → Consensus not reached
    - Validity testing was conducted at the data element and measure score level:
        - The developers rely on the measure's current use with CMS and its alignment with expert guidelines as demonstration of its face validity.
        - Face validity was not systematically assessed by recognized independent experts. Developers relied on its current use and alignment with national guidelines as proof of its face validity.
        - The developer uses approval by CMS and their contractors as evidence of validity of the measure. However, it was not clear whether it is the composite score of the individual component scores of the measures within the composite in the testing document submitted. They indicate consensus agreement that the use of diagnostic reference levels is a good indicator of quality and dose optimization, which was demonstrated by quotes from various organizations.
        - It appears they did not convene a panel to establish face validity but provided access to various reports. One SMP member wrote, “I’m not sure of the methods they used to collect, review and evaluate the literature they did review.”
        - There were mixed sentiments about whether the measure should be risk-adjusted. This comment, specifically, expressed those mixed sentiments: “The measure should be stratified by patient size, so each stratum is compared to size-based DRLs. This seems like a logical step.”
- **Ratings for construct composition:** H-2; M-3; L-0; I-1; → Measure passes with a MODERATE rating.
    - The developer demonstrated that performance on one of the component measures has little relationship on other measures, so each component measure does add something "new". The developer demonstrated that a weighted average (current measure) produces similar results to a straight average.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer (Appendix B)
- **Action items:**
    - The SMP needs to discuss and re-vote on the validity concerns.

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

Subgroup 1

#2860 30-Day, All-Cause, Unplanned Readmission Following Psychiatric Hospitalization in an Inpatient Psychiatric Facility (IPF)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This facility-level measure estimates an all-cause, unplanned, 30-day, risk-standardized readmission rate for adult Medicare fee-for-service (FFS) patients with a principal discharge diagnosis of a psychiatric disorder or dementia/Alzheimer's disease. The performance period for the measure is 24 months.
- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 49 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-0; M-8; L-1; 0-I; → Measure passes with MODERATE rating
    - Reliability testing was conducted at the measure score level:
        - The developers used two approaches to estimate the ICC: a split sample approach and an approach combining a split sample with bootstrapping. The split sample approach is likely to underestimate the ICC because it halves the sample size. The bootstrapping approach may overestimate the ICC if data is replaced after sampling.
            - For the split-half analysis, the ICC equaled 0.559. For the bootstrap method, ICC equaled 0.752.
            - The SMP members agreed that the methods were appropriate, and the results demonstrated good reliability of the measure.
- **Ratings for validity:** H-1; M-6; L-1; I-1; → Measure passes with MODERATE rating
    - Validity testing was conducted at the measure score level:
        - The developers used two approaches to test the measure's validity: construct validity was tested using Spearman rank order correlations and discriminant validity was tested using t-tests for between group differences of patient characteristics hypothesized to affect readmissions rates.
        - Construct validity was tested against the *Medication Continuation Following Inpatient Psychiatric Discharge* measure. The Spearman correlation was -0.300.
        - The SMP members found the methods and results acceptable for demonstrating construct validity.
        - Discriminant validity was tested against six patient characteristics hypothesized to be associated with higher readmissions rates: male patients, patients with a substance use disorder, patients with schizophrenia, non-White patients, patients with shorter length of stay at the IPF, and patients with socioeconomic characteristics associated with worse health outcomes.
        - Using t-tests to compare mean group differences, the findings were as hypothesized in four of the six areas.

        - Most SMP members found the methods and results acceptable for demonstrating discriminant validity.
    - SMP identified the following threats to validity:
        - Risk adjustment: Although the developer used a sound and thorough method to test the potential for risk adjustment and found that adjusting for social risk factors lead to reclassification of 4.6 percent of facilities, the developer decided not to risk-adjust. Some SMP members expressed concerns about this decision.
        - Meaningful differences: The Scientific Methods Panel members noted that the IQR is small and that the measure may only be able to identify outliers.

- Additional clarifying information from the developer (Appendix B)

Subgroup 2

#1598 Total Resource Use Population-Based PMPM Index

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The Resource Use Index (RUI) is a risk adjusted measure of the frequency and intensity of services utilized to manage a provider group's patients. Resource use includes all resources associated with treating members including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary and behavioral health services.
  A Resource Use Index when viewed together with the Total Cost of Care measure (NQF-endorsed #1604) provides a more complete picture of population based drivers of health care costs.
- **Type of measure:** Cost/Resource Use
- **Data source:** Claims
- **Level of analysis:** Clinician: Group/Practice, Population: Community, County or City
- **Risk-adjusted:** Statistical risk model; Other: this total resource use measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). Each of the 93 ACG actuarial cells can be considered a covariate of the multivariate risk model with the cell weights being the coefficients. The measure is also limited by insurance coverage to commercial only.
- **Sampling allowed**: None
- **Ratings for reliability:** H-4; M-3; L-0; I-2; → Measure passes with a HIGH rating
    - Reliability testing was conducted at the performance measure score level:
        - The developer used two methods to demonstrate the repeatability of the results, using bootstrapped averages with full replacement and a 90 percent random sampling without replacement approach, which approximates controlling for case mix.
        - The variances from Actual RUI ranged from -0.0037 to 0.0062 in the bootstrap to -0.0019 to 0.0016 in the 90 percent sample. The mean Total Resource Use results from the bootstrap and 90 percent samples compared to the actual RUI results for each provider group. The reliability testing demonstrates the repeatability of producing the same results a high proportion of the time.

        - One SMP member commented that both the testing methods, bootstrap and 90% random sampling, theoretically work for large sample. However, it is unclear how the results change when dealing with smaller providers (providers with less than 600 members). A couple of SMP members would like to see a signal-to-noise analysis.
- **Ratings for validity:** H-4; M-2; L-1; I-2; → Measure passes with a HIGH rating
    - Validity testing was conducted at both critical data element and performance measure score level:
        - Three approaches were used to assess validity: (1) critical data elements were correlated with each other and utilization, (2) performance measure score was validated against "known risk-adjusted utilization metrics", and (3) high/low performing groups were compared on utilization and RUI (although this largely shows the validity of the risk adjustor), and face validity was determined through a 45-day comment period.
        - Validity of Measure Components There is a high correlation between ACG score and the non-risk adjusted PMPM and Non-Risk Adjusted Total Cost Relative Resource Values (TCRRVs), which indicates that the non-risk adjusted PMPM and the non-risk adjusted TCRRVs are a good measure of resource use.
            - Correlations between the Non-Risk Adjusted Place of Service Metrics and Non-Risk Adjusted PMPMs & Non-Risk Adjusted TCRRVs are strong (ranging between 0.55 and 0.84).
            - The non-risk adjusted resource composite is highly correlated with ACGs, non-risk adjusted PMPMs, and non-risk adjusted TCRRVs (ranging between 0.77 and 0.95).
            - Professional RUI is correlated with overall RUI.
        - Validity of measure score:
            - The indexed Total Resource Use measure has a high correlation to a risk-adjusted composite utilization index, which was developed as a proxy to measure total resource consumption.
        - SMP reviewers generally agreed that the empirical validity tests are appropriate, the results are solid, and they are in the expected direction. However, one SMP member expressed interest in seeing the relationship of this measure to a quality construct.
        - A few of the reviewers commented on the use of ACG scores, stating that using a proprietary model limits the transparency and generalizability.
        - A couple of SMP reviewers also questioned the lack of race and social risk factors, especially about how income as a social risk factor was considered in the risk adjustment model.

#1604 Total Cost of Care Population-Based PMPM Index

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** Total Cost of Care reflects a mix of complicated factors such as patient illness burden, service utilization and negotiated prices. Total Cost Index (TCI) is a measure of a primary care provider's risk adjusted cost effectiveness at managing the population they care for. TCI

includes all costs associated with treating members including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary and behavioral health services.
A Total Cost Index when viewed together with the Total Resource Use measure (NQF-endorsed #1598) provides a more complete picture of population based drivers of health care costs.

- **Type of measure:** Cost/Resource Use
- **Data source:** Claims
- **Level of analysis:** Clinician: Group/Practice, Population: Community, County or City
- **Risk-adjusted:** Statistical risk model; Other: this total cost of care measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited by insurance coverage to commercial only. Each unique member is assigned one of 93 ACG actuarial cells, which has a corresponding weight that reflects relative illness burden (e.g., relative expected resource consumption). Attributed members are assigned a risk score based on diagnoses on claims from the performance measurement period, as well as member age and gender.
- **Sampling allowed**: None
- **Ratings for reliability:** H-4; M-3; L-0; I-2; → Measure passes with a HIGH rating
    - Reliability testing was conducted at the performance measure score level:
        - To measure the reliability of the TCI measure, the actual results were compared to the results calculated by two sampling methods, bootstrapping and a 90 percent random sample.
        - The differences between the actual TCI results and both the bootstrap and 90 percent sample results are very small, ranging from -0.0032 to 0.0066 in the bootstrap to -0.0026 to 0.0025 in the 90 percent sample.
        - Most SMP members agreed that the reliability tests are appropriate, and the results are good. One SMP member commented that the statistics of variance by provider can be better presented; one member questioned whether this measure is applicable for practices with low volume.
- **Ratings for validity:** H-4; M-2; L-1; I-2; → Measure passes with a HIGH rating
    - Validity testing was conducted at the data level:
        - Three approaches were used to assess validity: (1) critical data elements were correlated with each other and utilization, (2) performance measure score was validated against "known risk-adjusted utilization metrics", and (3) high/low performing groups were compared on utilization and RUI (although this largely shows the validity of the risk adjustor), and face validity was determined through a 45-day comment period.
        - Validity of Measure Components: There is a high correlation between the ACG score and the non-risk-adjusted PMPM and Non-Risk Adjusted Total Cost Relative Resource Values (TCRRVs).
            - Correlations between the Non-Risk-Adjusted Place of Service Metrics and Non-Risk-Adjusted PMPMs & Non-Risk-Adjusted TCRRVs are strong (ranging between 0.53 and 0.84).
            - The non-risk adjusted resource composite is highly correlated with ACGs, non-risk-adjusted PMPMs, and non-risk-adjusted TCRRVs (0.77 – 0.95)

        - Validity of measure score:
            - Both the overall price and total resource use are correlated with TCI as expected. However, price is more highly correlated with TCI because there is significantly more variation between providers in price than resource use; therefore it has a larger impact on TCI.
            - The indexed Total Cost of Care measure has a high correlation to a risk-adjusted composite utilization index, which was developed as a proxy to measure total resource consumption.
            - Providers' performance across all three measures is relatively consistent across all three years.
        - In general, the SMP reviewers agreed that there is a consistent picture across multiple correlations between measure components, and between the measure with relevant measures. However, one SMP member expressed interest in seeing the relationship of this measure to a quality construct.
        - A few of the reviewers commented on the use of ACG scores, stating that using a proprietary model limits the transparency and generalizability.
    - Several SMP members questioned the lack of race and SES indicators in the risk adjustment model.

#2431 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Acute Myocardial Infarction (AMI)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure estimates hospital-level, risk-standardized payment for an AMI episode-of-care starting with inpatient admission to a short term acute-care facility and extending 30 days post-admission for Medicare fee-for-service (FFS) patients who are 65 years of age or older with a principal discharge diagnosis of AMI.
- **Type of measure:** Cost/Resource Use
- **Data source:** Claims, Enrollment data
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 30 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** 3 high 5 moderate 1 low and 0 insufficient → Measure passes with a MODERATE rating
    - Reliability testing was conducted at the measure score level:
        - The developer calculated the intra-class correlation coefficient (ICC) using a split sample (i.e., test-retest) method and the Spearman-Brown prediction formula.
        - The ICC (for hospitals with 25 admissions or more), the agreement between the two independent assessments of the risk-standardized payment (RSP) for each hospital, is 0.681.
        - The SMP members generally agreed that the reliability test is appropriate, and the ICC result is moderate to high.
- **Ratings for validity:** H-1; M-5; L-2; I-0; → Measure passes with a MODERATE rating
    - Validity testing was conducted at the measure score level:

    - Face validity: Among the eight TEP members who responded to the developer's face validity question, three moderately agreed and five strongly agreed that this measure accomplished the purposes of measuring payments for Medicare patients for a 30-day AMI episode of care.
    - The developer calculated the correlation of the measure with the *Hospital Medicare Spending per Beneficiary (MSPB)* measure.
    - The AMI payment measure score was positively correlated with the Medicare Spending Per Beneficiary (MSPB) measure, with a correlation coefficient of 0.281 (p<.0001), meaning that higher spending across all Medicare FFS beneficiaries correlated with higher spending on patients hospitalized with AMI.
    - The observed payment breakdowns appropriately align with the distribution of the provider-level risk-standardized payments.
    - In general, the SMP members thought the face validity was fine, and the empirical test shows modest results and is in the right direction.
    - A couple of SMP members suggested ways to strengthen the analysis, (e.g., whether the hospital level spending was correlated with other measures of hospital spending in related services).
    - One member asked for clarification on the use of "ICD-9-to-CC assignment map".
- Additional clarifying information from the developer (Appendix B)

#2436 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Heart Failure (HF)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure estimates hospital-level, risk-standardized payment for a HF episode of care starting with inpatient admission to a short term acute-care facility and extending 30 days post-admission for Medicare fee-for-service (FFS) patients who are 65 years of age or older with a principal discharge diagnosis of HF.
- **Type of measure:** Cost/Resource Use
- **Data source:** Claims, Enrollment data
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 30 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-5; M-3; L-0; I-0; → Measure passes with a HIGH rating
    - Reliability testing was conducted at the measure score level:
        - The developer estimated the overall measure score reliability by calculating the intra-class correlation coefficient (ICC) using a split sample (i.e., test-retest) method.
        - They calculated the ICC for hospitals with 25 admissions or more. The agreement between the two independent assessments of the risk-standardized payment (RSP) for each hospital is 0.781.
        - The SMP members generally agreed the reliability test is appropriate and the result shows high reliability.
- **Ratings for validity:** H-2; M-4; L-2; I-0; → Measure passes with a MODERATE rating

- Validity testing was conducted at the performance measure score level:
  - Face validity: Among the 8 TEP members who provided a response, one responded "Somewhat Agree," three responded "Moderately Agree," and four reported "Strongly Agree" that this measure accomplished the purposes of measuring payments for Medicare patients for a 30-day HF episode of care.
  - The developer assessed the measure's correlation with the *Hospital Medicare Spending per Beneficiary (MSPB)* measure.
  - The HF payment measure score was positively correlated with the Medicare Spending Per Beneficiary (MSPB) measure with a correlation coefficient of 0.543, meaning that higher spending across all Medicare FFS beneficiaries correlated with higher spending on patients hospitalized with HF.
  - The observed payment breakdowns appropriately align with the distribution of the provider-level risk-standardized payments.
  - The SMP members generally agreed the validity test is appropriate and the result is strong.
  - One member asked for clarification of the use of "ICD-9-to-CC assignment map".
  - One member noted that the distribution of measure scores across hospitals shows that there are meaningful differences, although the distribution is fairly tight.
  - There is also concern regarding a large difference in mean cost with and without the social risk factors (dual status and low SES).
- Additional clarifying information from the developer (Appendix B)

#2579 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Pneumonia (PN)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure estimates hospital-level, risk-standardized payment for an eligible pneumonia episode of care starting with inpatient admission to a short term acute-care facility and extending 30 days post-admission for Medicare fee-for-service (FFS) patients who are 65 years or older with a principal discharge diagnosis of pneumonia or principal discharge diagnosis of sepsis (not including severe sepsis) that have a secondary discharge diagnosis of pneumonia coded as present on admission (POA) and no secondary diagnosis of severe sepsis coded as POA.
- **Type of measure:** Cost/Resource Use
- **Data source:** Claims, Enrollment data
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 57 risk factors
- **Sampling allowed**: None
- **Ratings for reliability:** H-5; M-3; L-0; I-0; → Measure passes with a HIGH rating.
  - Reliability testing was conducted at the measure score level:
    - The developer estimated the overall measure score reliability by calculating the intra-class correlation coefficient (ICC) using a split sample (i.e., test-retest) method.

        - They calculated the ICC for hospitals with 25 admissions or more. The agreement between the two independent assessments of the risk-standardized payment (RSP) for each hospital is 0.815.
        - The SMP members generally agreed the reliability test is appropriate and the result shows high reliability.
- **Ratings for validity:** H-2; M-4; L-2; I-0; → Measure passes with a MODERATE rating
    - Validity testing was conducted at the measure score level:
        - Face validity: Among the 10 TEP members who provided a response, one responded "Somewhat Agree," three responded "Moderately Agree," and six reported "Strongly Agree" that this measure accomplished the purposes of measuring payments for Medicare patients for a 30-day pneumonia episode of care.
        - The developer assessed the measure's correlation with the hospital Medicare Spending per Beneficiary (MSPB) measure.
        - The pneumonia payment measure score was positively correlated with the Medicare Spending Per Beneficiary (MSPB) measure with a correlation coefficient of 0.588 ($p<0.0001$), meaning that higher spending across all Medicare FFS beneficiaries correlated with higher spending on patients hospitalized with pneumonia.
        - The observed payment breakdowns appropriately align with the distribution of the provider-level risk-standardized payments.
        - The SMP members generally agreed the validity test is appropriate and the result is strong.
        - A couple of SMP members voiced concerns over the decision to not include dual eligibility as a risk factor despite the evidence showing its significant association with payments.
- Additional clarifying information from the developer (Appendix B)

#3610 30-Day Risk-Standardized Morbidity and Mortality Composite Following Transcatheter Aortic Valve Replacement (TAVR) (American College of Cardiology)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The TAVR 30-day morbidity/mortality composite is a hierarchical, multiple outcome risk model that estimates risk standardized results (reported as a "site difference") for the purpose of benchmarking site performance. This measure estimates hospital risk standardized site difference for 5 endpoints (death from all causes, stroke, major or life-threatening bleeding, acute kidney injury, moderate or severe paravalvular aortic regurgitation) within 30 days following transcatheter aortic valve replacement. The measure uses clinical data available in the STS/ACC TVT Registry for risk adjustment for the purposes of benchmarking site to site performance on a rolling 3-year timeframe.
- **Type of measure:** Composite
- **Data source:** Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 43 variables.
- **Sampling allowed**: No

- **Ratings for reliability:** H-0; M-7; L-1; I-0; → Measure passes with MODERATE rating
    - Reliability testing was conducted at the measure score level:
        - The developer estimated hospital-specific performance using a hierarchical proportional odds model on 100 sets of simulated data. Then, they calculated the Pearson correlation coefficient between each hospital's calculated estimate and the simulated true value. Reliability was calculated as the average squared Pearson correlation coefficient across the 100 data sets.
        - The overall estimated reliability was 0.64, with a range from 0.65 for hospitals with at least 25 cases (n = 278) to 0.73 for hospitals with at least 200 cases (n = 96). The developer indicates they will be using a minimum of 60 cases over a three-year period for public reporting.
        - In general, SMP subgroup members found the testing methodology appropriate and that the results supported moderate reliability.
- **Ratings for validity:** H-3; M-5; L-0; I-0; → Measure passes with MODERATE rating
    - Validity testing was conducted at the composite measure score and component measure score level:
        - The developer assessed the validity of the composite measure score using a known-group analysis. They divided the facilities into three levels of performance based on the global rank composite (i.e., better than expected, as expected, and worse than expected). Then, they examined the adjusted observed to expected (O/E) odds ratios for the individual components for each group. Sites with better than expected performance on the global rank composite metric showed lower O/E ratios when compared with sites that performed as expected or worse than expected. Sites that performed worse than expected showed consistently higher O/E ratios than other sites.
        - The developer assessed the validity of the component measure scores using Cox proportional hazards modeling to evaluate the associations of the components with one-year mortality and average change in KCCQ-OS. All four non-fatal complications (components) were found to be associated with increased risk of one-year mortality and patient-reported health status (assessed via KCCQ-OS score).
        - Exclusion of hospitals with more than 10 percent missing data for the global rank endpoint, baseline Kansas City Cardiomyopathy Questionnaire 12 (KCCQ-12) or baseline 5-meter walk test resulted in the exclusion of over half of the hospitals in the initial cohort (59,904 of 114,121).
        - Covariates for case-mix adjustment were pre-selected based on inclusion in the risk model for NQF #3534 (TAVR 30-day mortality). Covariates were retained in the model regardless of their statistical significance. The developer did not collect or analyze any variables that directly measure social risk, based on the social risk analysis conducted for NQF #3534.
        - Table 1 : C-statistic for Predicting an Outcome in One of the Worst Ranking Categories

Rank ≤ 1	Rank ≤ 2	Rank ≤ 3	Rank ≤ 4	Rank ≤ 5

0.70	0.65	0.63	0.64	0.63

  - The SMP subgroup members felt that the associations demonstrated through the analysis supported moderate to high validity.
- **Ratings for composite construction:** H-3; M-3; L-1; I-1; → Measure passes with MODERATE rating
  - Composite construction:
    - The global ranking endpoint is an ordinal categorical variable having six levels in which category one represents the worst possible outcome (death) and category six represents the best possible outcome (alive and free of major complications). Patients are classified according to the worst outcome (lowest rank score) that they experience. Endpoints were ranked in order of their decreasing hazard ratios with one-year mortality.
    - The clinical importance of the complications was confirmed by assessing their associations with one-year mortality and one-year KCCQ-OS.
    - The SMP sub-group members generally supported the composite construction. A couple of members questioned whether this measure represents a composite measure or a composite outcome and whether the additional complexity of this approach resulted in more precise measurement.
- The SMP did not have any substantial concerns regarding the scientific acceptability of this measure.

#3623 Elective Primary Hip Arthroplasty

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The Elective Primary Hip Arthroplasty episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who receive an elective primary hip arthroplasty during the performance period. The measure score is a clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from the 30 days prior to the clinical event that opens or "triggers" the episode, through 90 days after the trigger. Patient populations eligible for the Elective Primary Hip Arthroplasty measure include Medicare beneficiaries enrolled in Medicare Parts A and B. The Elective Primary Hip Arthroplasty measure will be reported for TINs and TIN-NPIs with 10 or more episodes. The measure is used in the Merit-based Incentive Payment System (MIPS) for MIPS performance period 2020 onwards.
- **Type of measure:** Cost/Resource Use
- **Data source:** Claims
- **Level of analysis:** Clinician: Group/Practice, Clinician: Individual
- **Risk-adjusted:** Statistical risk model with 121 risk factors, including patient health status and clinical factors.
- **Sampling allowed**: N/A
- **Ratings for reliability:** H-7; M-1; L-0; I-0; → Measure passes with a HIGH rating
  - Reliability testing was conducted at the performance measure score level:
    - The developer conducted both signal-to-noise and split sample tests for measure reliability.

        - At a volume threshold of at least 10 episodes, the signal-to-noise test found that the mean reliability is 0.86 for TINs and 0.80 for TIN-NPIs. When examined by a number of clinicians within the practice, the average reliability score increases from 0.78 (1 clinician) to 0.98 (21+ clinicians) for TINs.
        - The split-sample testing found that the ICC coefficient was 0.78 at the TIN-level and 0.73 at the TIN-NPI level.
        - The SMP members agreed that both approaches are appropriate, and the testing results indicated high measure score reliability.
- **Ratings for validity:** H-1; M-5; L-2; I-0; → Measure passes with MODERATE rating
    - Validity testing was conducted at the performance measure score level:
        - The developer conducted both empirical validity testing and a systematic assessment of face validity through a TEP.
        - Out of the 11 TEP respondents to the measure face validity survey, all 11 (100 percent) agreed that each of the measure specifications helps the measure capture clinician cost performance as intended and the scores from the measure as currently specified provide an accurate reflection of clinician cost effectiveness.
        - The developer also evaluated the empirical validity of this measure by examining correlation with an NQF-endorsed measure of resource use: the *Medicare Spending per Beneficiary (MSPB) Hospital* measure (NQF# 2158), which assesses the risk-adjusted cost to Medicare for services performed by hospitals and other healthcare providers during an MSPB-Hospital episode. Provider cost scores for the Hip Arthroplasty Measure decrease (i.e., cost performance improves) with increases in MSPB Hospital Measure performance ratings. As MSPB performance ratings increase (from 0 to 5-10), the O/E ratios in the Mean column decrease from 1.08 to 0.97, indicating a consistent improvement in Hip Arthroplasty Measure performance as well.
        - In general, the SMP members agreed that the face validity was assessed in a robust way, but they raised some questions regarding the empirical testing:
            - It is not clear how the MSPB hospital rating was attributed to TINs or TIN-NPIs.
            - The correlation between this measure and MSPB is okay but not conceptually compelling.
            - One member raised concerns about including DRG 469 and 470 in the risk adjustment model.
            - A couple of members noted that social risk factors have limited marginal effect on scores, but the actual effect is not reported.
            - There is a concern about excluding patients who may potentially die during the episode duration due to low-quality treatment.

#3625 Non-Emergent Coronary Artery Bypass Graft (CABG)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The Non-Emergent CABG episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo a CABG procedure during the performance

period. The measure score is the clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from 30 days prior to the clinical event that opens, or "triggers," the episode through 90 days after the trigger. Patient populations eligible for the Non-Emergent CABG measure include Medicare beneficiaries enrolled in Medicare Parts A and B. The Non-Emergent CABG measure will be reported for TINs and TIN-NPIs with 10 or more episodes. The measure is used in the Merit-based Incentive Payment System (MIPS) for MIPS performance period 2020 onwards.

- **Type of measure:** Cost/Resource Use
- **Data source:** Claims
- **Level of analysis:** Clinician: Group/Practice, Clinician: Individual
- **Risk-adjusted:** Statistical risk model with 119 risk factors, including patient health status and clinical factors.
- **Sampling allowed**: N/A
- **Ratings for reliability:** H-4; M-4; L-0; I-0; → Measure passes with MODERATE rating
    - Reliability testing was conducted at the performance measure score level:
        - The developer conducted both signal-to-noise and split sample tests for measure reliability.
        - At a volume threshold of at least 10 episodes, the mean reliability for TINs is 0.84 and for TIN-NPIs is 0.75. When examined by the number of clinicians within the practice, the average reliability scores increased from 0.76 (1 clinician) to 0.97 (21+ clinicians) for TINs.
        - The split-sample testing found that the ICC coefficient was 0.80 at the TIN-level and 0.64 at the TIN-NPI level.
        - The SMP members agreed that both approaches are appropriate, and the testing results indicated high measure score reliability.
- **Ratings for validity:** H-0; M-5; L-3; I-0; → Measure passes with MODERATE rating
    - Validity testing was conducted at the performance measure score level:
        - The developer conducted both empirical validity testing and a systematic assessment of face validity through a TEP.
            - Out of the nine TEP respondents to the survey, all nine (100 percent) agreed that each of the measure specifications helps the measure capture clinician cost performance as intended, and eight (89 percent) agreed that the scores from the measure, as currently specified, provide an accurate reflection of clinician cost effectiveness.
            - Provider cost scores for the Non-Emergent CABG Measure decrease (i.e., cost performance improves) with increases in MSPB Hospital Measure performance ratings. As MSPB performance ratings increase (from 0 to 5-10), the O/E ratios in the Mean column decrease from 1.03 to 0.98, indicating a consistent improvement in Non-Emergent CABG Measure performance as well.
        - In general, the SMP members agreed that the face validity was assessed in a robust way, but they raised some questions regarding the empirical testing:
            - There is not enough variability between the SMPB performance rating categories to show meaningful information.

- They would like to see clarification of how costs are narrowly defined to only be those associated with Lumbar Fusion and its post-acute care.
- Exclusion: It is a major concern that almost half the episodes are excluded at both levels (TIN and TIN-NPI); another concern pertains to excluding patients who may potentially die during the episode duration due to low-quality treatment.
- Several SMP members raised concerns about the risk adjustment model. There were questions about not including social risk factors, especially dual eligibility status, in the final risk adjustment model and the clustering of factors associated with homebound status and when they constitute unmanageable clinical risk.

#3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels episode-based cost measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo surgery for lumbar spine fusion during the performance period. The measure score is the clinician's risk-adjusted cost for the episode group averaged across all episodes attributed to the clinician. This procedural measure includes costs of services that are clinically related to the attributed clinician's role in managing care during each episode from 30 days prior to the clinical event that opens, or "triggers," the episode through 90 days after the trigger. Patient populations eligible for Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure include Medicare beneficiaries enrolled in Medicare Parts A and B. The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure will be reported for TINs and TIN-NPIs with 10 or more episodes. The measure is used in the Merit-based Incentive Payment System (MIPS) for MIPS performance period 2020 onwards.
- **Type of measure:** Cost/Resource Use
- **Data source:** Claims
- **Level of analysis:** Clinician: Group/Practice, Clinician: Individual
- **Risk-adjusted:** Statistical risk model with 122 risk factors, including patient health status and clinical factors. The Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels measure is stratified into three sub-groups, or mutually exclusive and exhaustive divisions of the overall episode group: one-level lumbar fusion, two-level lumbar fusion, and three-level lumbar fusion
- **Sampling allowed**: N/A
- **Ratings for reliability:** H-4; M-4; L-0; I-0; → Measure passes with MODERATE rating
    - Reliability testing was conducted at the performance measure score level:
        - The developer conducted both signal-to-noise and split sample tests for measure reliability.
        - At a volume threshold of at least 10 episodes, the mean reliability for TINs is 0.78 and 0.72 for TIN-NPIs. When examined by a number of clinicians within the practice, the average reliability scores increased from 0.71 (1 clinician) to 0.95 (21+ clinicians) for TINs.
        - The split-sample testing found that the ICC coefficient was 0.73 at the TIN-level and 0.67 at the TIN-NPI level.

        - The SMP members agreed that both approaches are appropriate and the testing results indicated high measure score reliability. One SMP member asked about clarification on the use of a 10-episode testing volume threshold.
- **Ratings for validity:** H-0; M-6; L-2; I-0; → Measure passes with a MODERATE rating
    - Validity testing was conducted at the performance measure score level:
        - The developer conducted both empirical validity testing and a systematic assessment of face validity through a TEP.
            - Out of the nine TEP respondents to the survey, substantial majorities (six to eight respondents) agreed that each of the measure specifications helps the measure capture clinician cost performance as intended, and the scores from the measure, as currently specified, provide an accurate reflection of clinician cost effectiveness.
        - Provider cost scores for the Lumbar Spine Fusion Measure decrease (i.e., cost performance improves) with increases in MSPB Hospital Measure performance ratings. As MSPB performance ratings increase (from 0 to 5-10), the O/E ratios in the Mean column decrease from 1.04 to 0.96, indicating a consistent improvement in Lumbar Spine Fusion Measure performance as well.
        - In general, the SMP members agreed that the face validity was assessed in a robust way, but they raised some questions regarding the empirical testing:
            - There is not enough variability between the SMPB performance rating categories to show meaningful information.
        - Additional aspects of validity (e.g., predictive) should be tested over time.
            - Exclusion: There is a concern regarding the exclusion of patients who may potentially die during the episode duration due to low-quality treatment.
            - They would like to see clarification of how costs are narrowly defined to only be those associated with Lumbar Fusion and its post-acute care.
            - A couple of members raised concerns about not including social risk factors, especially dual status, in the final risk adjustment model.
- Additional clarifying information from the developer (Appendix B)

Subgroup 3

#2902 Contraceptive Care - Postpartum

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** Among women ages 15 through 44 who had a live birth, the percentage that is provided: 1) A most effective (i.e., sterilization, implants, intrauterine devices or systems (IUD/IUS)) or moderately (i.e., injectables, oral pills, patch, or ring) effective method of contraception within 3 and 60 days of delivery. 2) A long-acting reversible method of contraception (LARC) within 3 and 60 days of delivery.
  Two time periods are proposed (i.e., within 3 and within 60 days of delivery) because each reflects important clinical recommendations from the Centers for Disease Control and Prevention (CDC) and the American College of Obstetricians and Gynecologists (ACOG). The 60-day period reflects ACOG recommendations that women should receive contraceptive care at

the 6-week postpartum visit. The 3-day period reflects CDC and ACOG recommendations that the immediate postpartum period (i.e., at delivery, while the woman is in the hospital) is a safe time to provide contraception, which may offer greater convenience to the client and avoid missed opportunities to provide contraceptive care.

- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims
- **Level of analysis:** Clinician: Group/Practice, Facility, Health Plan, Population: Regional and State
- **Not risk-adjusted**
- **Sampling allowed**: None.
- **Ratings for reliability:** H-2; M-6; L-0; I-0; → Measure passes with MODERATE rating
    - Reliability testing was conducted at the measure score level:
        - The developer calculated signal-to-noise for all three levels of the measure specification: (1) group/practice, (2) health plan, and (3) public health region using a beta-binomial model using parametric empirical Bayes methods, which is appropriate for the measure. Targets greater than 0.90 may be used for high-stake purposes and greater than 0.70 used for reporting and monitoring with claims data sourced from Iowa Medicaid, CMS data set for Texas, Washington State Health Care Authority, Massachusetts Health Dataset, and Louisiana Medicaid Dataset.
        - Results are generally moderate, depending upon the region (state) being discussed, the type of contraception, and age group. Unit sizes greater than 75 improved group billing provider levels to greater than 0.80.
        - Results are not "pooled" across all geographical regions, although testing seeks reliability estimates for different regions, levels of reporting, contraception method, and age groups.
        - Beyond age, state, and program, no other disparities data are provided. Evidence also demonstrates health disparities for six-week postpartum visits, which impacts performance for providing the 60-day method of contraception.
- **Ratings for validity:** H-0; M-5; L-3; I-0; → Measure passes with MODERATE rating
    - Validity testing was conducted at the measure score level:
        - The face validity of nine experts was conducted to distinguish good versus poor quality care: MOST/MOD (mean 4.22, median 4.5), LARC (mean 3.78, median 4).
        - The developer performed construct validity testing of the measure to (1) timeliness of prenatal care, and (2) postpartum care measures. Pearson correlations and a novel multilevel correlation estimation method (due to low volume events in high volume populations) were used in greater than or equal to 25 patients.
        - Pearson correlations and multilevel correlation estimates to the two other quality measures were MOST/MOD 3 day (0.21-0.31, 0 .37-0.39), MOST/MOD 60-day (0.28-0.52, 0.52-0.60), LARC 3-day (0.06-0.28, 0.10 to 0.32), and for LARC 60-day (0.30-0.45, 0.42 to 0.51). Pearson was lower than the multilevel estimation with many showing weak correlations to moderately positive.
        - A reviewer noted the measure might assume what patients want, may want, or the timing of decision making inappropriately, which could compromise validity.

    - Empirical validity testing was not conducted for health plans and populations. The developer noted this was due to the limited numbers of units (n≤21) at these levels, which are not sufficient for correlation testing.
    - The measure is not risk-adjusted, yet it is stratified by adolescents and adults. No testing was provided to support the stratification approach.
- Additional clarifying information from the developer (Appendix B)

#2903 Contraceptive Care – Most & Moderately Effective Methods

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The percentage of women aged 15-44 years at risk of unintended pregnancy that is provided a most effective (i.e., sterilization, implants, intrauterine devices or systems (IUD/IUS)) or moderately effective (i.e., injectables, oral pills, patch, or ring) method of contraception. The measure is an intermediate outcome measure because it represents a decision that is made at the end of a clinical encounter about the type of contraceptive method a woman will use, and because of the strong association between type of contraceptive method used and risk of unintended pregnancy.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims
- **Level of analysis:** Clinician: Group/Practice, Facility, Health Plan, Population: Regional and State
- **Not risk-adjusted**
- **Sampling allowed**: None.
- **Ratings for reliability:** H-5; M-3; L-0; I-0; → Measure passes with HIGH rating
    - Reliability testing was conducted at the measure score level. Data element validity testing was conducted; therefore, additional data element reliability testing is not required.
    - The measure level of analysis includes the following levels: Clinician: Group/Practice, Facility, Health Plan, Population: Regional and State. Reliability testing is provided in state-level payer programs, although not all-payer state programming.
    - Several reviewers had concerns regarding performance not being measured in the last two months of the year and could disincentivize positive performance.
    - Using the beta-binomial model and the parametric empirical Bayes methods (which is appropriate for the measure), measure score reliability was calculated in signal-to-noise analyses for all four levels: Clinician: Group/Practice, Facility, Health Plan, Population: Regional and State.
    - Claims data from seven organizations were utilized for testing: Iowa Medicaid Enterprise (2018), Iowa Department of Public Health (IDPH) (2019), New York Presbyterian Hospital/Columbia University Irving Medical Center (2018), Washington State Health Care Authority (2019), Massachusetts Mass Health (2019), Oregon Medicaid (2015) and Louisiana Medicaid Program (2019).
    - Planned Parenthood Federation of America (2019) and Title X Family Planning Program (2019) were also included using different calculations and interpretations as the patient population is women seeking reproductive care.
    - Reliability scores were very high at all testing levels, except the group level. Many reviewers prefer case limits, such as the 75 case counts obtained at group level,

especially in high stakes program use. Targets greater than 0.90 may be used for high-stake purposes and greater than 0.70 used for reporting and monitoring. The developer emphasizes the measure is not to be used in pay for performance programs.

- **Ratings for validity:** H-1; M-5; L-2; I-0; → Measure passes with MODERATE rating
  - Validity testing was conducted at the measures score and data element levels:
    - Measure score validity testing was not conducted for health plans as populations as the limited numbers of units for these levels were not sufficient for correlation testing.
    - The developer performed construct validity testing of the measure to (1) Cervical Cancer Screening, (2) Chlamydia Screening, (3) Encounter for Contraceptive Counseling, and (4) Encounter for Gynecological Exam Measures, hypothesizing measured entities performing well on contraceptive care should perform well on the other measures, and stated the correlation magnitude may be weak for cervical cancer screening and chlamydia screening with screening frequency differences.
    - Pearson correlations and a novel multilevel correlation estimation method (due to low volume events in high volume populations) were used with thresholds of 25, 50, and 75 eligible patients. The novel approach generally showed slightly higher or similar correlations to Pearson's for Contraceptive Counseling and Gynecological Examination measures in group reporting with moderate reliability. The Cervical Cancer Screening and Chlamydia Screening measures generally showed slightly higher or the same correlations to Pearson's than the novel approach, except 21-44 in Chlamydia Screening. The submitted measure showed "just" to poor reliability for these two measures. As predicted, the correlations were weak to none in the Planned Parenthood Federation of America in Cervical Cancer Screening and Chlamydia Screening measures possibly due to screening frequency differences.
    - Data element validity testing was conducted with 423 patients, compared claims vs. patient record for 10 critical data elements in calculated sensitivity, specificity, PPV, NPV, Cohen's Kappa statistics with 95 percent CIs, and percent agreement for each data element. Sensitivity was above 0.5 for most data elements, except the contraceptive patch, in which specificity, PPV, and NPV were greater than 0.8 for all data elements. Percent agreement was greater than 80 percent for all data elements.
    - Reviewers were concerned about sensitivity results being "less than desirable," specifically with the contraceptive patch of 0.25 used to define numerator.
    - Face validity was conducted with nine independent panel experts to assess whether the measure will reflect quality of contraceptive care. The mean rating measure was 4.67 with a median of 5 (Strongly Agree), range 4-5. One reviewer was "unclear on patient-centeredness of this overall (face validity)".
    - The measure is not risk-adjusted, yet it is stratified by adolescents and adults. Multiple reviewers had concern with the lack of social risk stratification. The developer stated, "statistically significant differences by age group (for ages 20-

29 compared to ages 30-44) and among women who have never been married (compared to women of other marital status", were identified, yet "no significant differences occur between race/ethnicity, most categories of marital status, and poverty level" were seen. These findings contrast the identified disparities from measure #2902 with overlapping populations.

- Additional clarifying information from the developer (Appendix B)

#2904 Contraceptive Care - Access to LARC

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** Percentage of women aged 15-44 years at risk of unintended pregnancy that is provided a long-acting reversible method of contraception (i.e., implants, intrauterine devices or systems (IUD/IUS)). It is an access measure because it is intended to identify very low rates (less than 1-2%) of long-acting reversible methods of contraception (LARC), which may signal barriers to LARC provision.
- **Type of measure:** Structure
- **Data source:** Claims
- **Level of analysis:** Clinician: Group/Practice, Facility, Health Plan, Population: Regional and State
- **Not risk-adjusted**
- **Sampling allowed**: None.
- **Ratings for reliability:** H-3; M-5; L-0; I-0; → Measure passes with MODERATE rating
    - Reliability testing was conducted at the measure score level. Data element validity testing was conducted; therefore, additional data element reliability testing is not required.
    - Several reviewers had concerns about performance not being measured in the last two months of the year, stating that it contradicts the premise of measure #2902.
    - The denominator population in #2904, "deliveries that did not end in a live birth (i.e., miscarriage, ectopic, stillbirth or induced abortion," is excluded in #2902. This includes a population previously excluded.
    - Identifying enrollment gaps and Section 1115 waivers in claims may be difficult.
    - Claims data is sourced from six organizations for testing: Iowa Medicaid Enterprise (2018), Iowa Department of Public Health (IDPH) (2019), Title X Grantee (New York Presbyterian Hospital/Columbia University Irving Medical Center (2018), Washington State Health Care Authority (2019), Massachusetts Mass Health (2019), and Louisiana Medicaid Program (2019).
    - Planned Parenthood Federation of America (2019) and Title X Family Planning Program (2019) were included with different calculations and interpretations as the population is women seeking care from reproductive health clinics.
    - The developer calculated the measure score reliability signal-to-noise for all four levels of the measure specification: (1) Clinician: Group/Practice, (2) Facility, (3) Health Plan, Population: Regional, and (4) State using a beta-binomial model and parametric empirical Bayes methods, which is appropriate for the measure.
    - Targets greater than 0.90 may be used for high-stake purposes and greater than 0.70 used for reporting and monitoring. The developer emphasizes the measure is not to be used in pay for performance programs.

- Reliability results were greater than 0.70 at the facility and health plan levels and consistently greater than 0.90 at the public health region level. At the group level, estimates were greater than 0.70 if the measure is restricted to practices with greater than 75 patients.
- Multiple reviewers were concerned with low sensitivity for the live birth data element used to establish an exclusion criterion: 2019 WA HCA 0.9 percent, 2018 IME 4.6 percent, 2019 PPFA 0 percent and may identify poor data element reliability.
- Some reviewers expressed the following concern: "if a reporting entity has no or very few women using LARC (e.g., less than 2%), barriers restricting LARC access might be present and should be investigated." No analysis was conducted on the reliability of being classified as a low outlier.

- **Ratings for validity:** H-0; M-7; L-1; I-0; → Measure passes with MODERATE rating
  - Validity testing was conducted at the measure score and data element levels:
    - Construct validity testing of the measure was conducted to (1) Cervical Cancer Screening, (2) Chlamydia Screening, (3) Encounter for Contraceptive Counseling, and (4) Encounter for Gynecological Exam Measures, hypothesizing measured entities performing well on contraceptive care should perform well on the other measures. They stated the correlation magnitude may be weak for cervical cancer screening and chlamydia screening with screening frequency differences.
    - Pearson correlations and a novel multilevel correlation estimation method (due to low-volume events in high volume populations) were used with thresholds of 25, 50, and 75 eligible patients. For greater than or equal to 75 patients in all testing, the facility level Pearson's ranged from 0.23 to 0.78 across all age groups, the highest with Gynecological Examination. In the novel approach, the range in groups was 0.78 to 0.98. At the group level, Pearson's ranged from 0.08 to 0.67 across all age groups, the highest with Contraceptive Counseling. In the novel approach, the range across all groups was 0.06 to 0.67.
    - The Planned Parenthood Federation of America measure score validity testing to the four measures ranges and ages: Pearson's (0.23-0.78), Novel (0.78-0.99), with a 95 percent CI (0.66-0.99).
    - Critical data element validity testing was conducted with 423 patients. It compared claims versus patient record for 10 critical data elements in a calculated sensitivity greater than 0.5 for most of the data elements (except contraceptive patch); specificity, PPV, and NPV were greater than 0.8 for all data elements, Cohen's Kappa statistics with 95 percent CIs ranged from 0.567 to 1.000, and percent agreement greater than 80 percent for each data element (two LARC methods and three exclusion criteria elements).
    - Face validity was conducted by nine expert panelists to rate the measure on its ability to discern good from poor quality care. The mean rating for this measure was 4.33 with a median of 4.5 (between Agree and Strongly Agree), range 3-5.
    - The measure is not risk-adjusted. The developers cited a 2015-2017 National Survey of Family Growth (NSFG), stating no significant differences exist by age group, race/ethnicity, marital status, and poverty level. The measure was not stratified based on social risks.

- Additional clarifying information from the developer (Appendix B)

3501e: Hospital Harm – Opioid-Related Adverse Events

MEASURE HIGHLIGHTS

- This is a new measure. It was previously discussed by SMP and passed. However, it failed with Standing Committee because there were clinical concerns surrounding how the measure was defined and constructed.
- **Description:** This measure assesses the proportion of inpatient hospital encounters where patients ages 18 years of age or older have been administered an opioid medication, subsequently suffer the harm of an opioid-related adverse event, and are administered an opioid antagonist (naloxone) within 12 hours. This measure excludes opioid antagonist (naloxone) administration occurring in the operating room setting.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Records
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: None.
- **Ratings for reliability:** H-2; M-5; L-0; I-1; → Measure passes with a MODERATE rating
    - Reliability testing was conducted at the data element level:
        - For data element reliability, the developer compared electronically extracted data to manually abstracted data using kappa to quantify agreement. The kappa was 0.98 at one site and 1.00 at all other sites for the six randomly selected sub-samples, comparing the electronically extracted EHR data to manually extracted EHR data for the same medical record.
- **Ratings for validity:** H-1; M-6; L-1; I;0 → Measure passes with a MODERATE rating.
    - Validity testing was conducted at the data element and measure score level:
        - Data element: Tested inter-rater agreement by comparing the hospitals' EHR data to a clinical abstractor (as described above in the reliability section). The agreement rate between data electronically extracted from the sampled patients' EHR and data manually abstracted from the medical records was 100 percent for all but two data elements. Measure score validity was assessed for this rather small sample by PPV, sensitivity, NPV, and specificity. PPV was 100 percent, and sensitivity is 100 percent in all but one test site. NPV is also 100 percent. Specificity is 100 percent.
        - Score-level: An EHR-reported opioid related adverse event was compared to clinical review of the patient's chart. Given the strict definition of the numerator event, this does not appear to be much different than validating the data elements that comprise the measure.
        - There was a concern that the measure score was calculated and tested at the patient-level, not at the entity level.
        - One SMP member was concerned with the lack of risk adjustment, given that there is drug diversion in hospitals and some patients have conditions or procedures that make safe opioid administration more difficult.

- Additional clarifying information from the developer (Appendix B)

Appendix B: Additional Information Submitted by Developers for Consideration

Subgroup 1

Measure Number: 2880

Measure Title: Excess days in acute care (EDAC) after hospitalization for heart failure

Measure Developer: Yale/CORE; Steward: CMS

We thank Scientific Methods Panel members for their thoughtful comments and for the time they took to review the measure specifications and testing results. Below we provide responses to questions posed by the Panel members within their Preliminary Assessment.

Reliability

- **Issue 1:** One Panel member asked about the impact of excluding heart failure admissions within 30 days of a prior heart failure admission.
    - **Developer Response 1:** Admissions for a condition within 30 days of discharge from an index admission for that same condition are excluded as index admissions. Thus, no hospitalization will be considered as both a readmission and an index admission within the same measure. For this measure, this excludes 112,210 of a total of 1,286,352 admissions, or about 8%. This exclusion is aligned with exclusions in the respective readmission measures.

Validity

- **Issue 1:** Panel members expressed concern regarding the validity testing because the approach examined correlated of the EDAC measures with other measures that include the same readmission events, without correction for the overlap. Another Panel member indicated that process-outcome correlations or pre-post analyses of intervention effects are strongly preferred.
    - **Developer Response 1:** As noted in our testing attachment, developers often do not have access to the type of data that would ideally be used for the purposes of empiric validity testing, such as patient-level data on process measures that are related to the outcome.
    - We have, however, included updated testing results that implement the methodologic approach suggested by the Scientific Methods Panel with respect to Star Ratings (removing the comparator measure from the Star Rating Readmission Group score before analyzing the correlation). In addition, because conceptually there could be a relationship between EDAC and the other domains of Star Ratings (which include measures related to care coordination in the domain of Patient Experience, hospital-acquired infections in the Safety of Care domain, in addition to measures related to timeliness and effective care in the Emergency Department, as well as mortality) we also examined the association between the EDAC measure and Star Ratings after removing the entire Readmission Group score. Finally, we separately examined the relationship

between HF EDAC and components of the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, a validated and NQF-endorsed patient-reported quality measure that reflects processes of care that relate, conceptually, to post-discharge hospital visits.

- Our results show that there remains a moderate correlation in the expected direction between the EDAC measure and Star Ratings even after removing the overlapping measure (Table 1A). We also find an association, albeit weaker, between the EDAC measure and Star Ratings after removing the entire Readmission Group from Star Ratings (Table 1B), suggesting that the EDAC measures share a quality signal with the remaining Star Ratings domains, which include measures related to patient safety, timeliness and effective care, patient experience, and mortality.
- Consistent with the results seen for Star Ratings, our results show an association in the hypothesized direction with components of HCAHPS (Table 2). Taken together, the results provide additional support for the validity of the HF EDAC measure. The results are discussed in more detail below.
- Relationship with Star Ratings
- Briefly, using the recently **updated Star Ratings methodology** that no longer uses the latent variable model, but rather averages the performance of measures in the measure group, we compared hospitals' performances on HF EDAC with performance on Star Ratings summary scores and Readmission Group scores, with and without the HF EDAC measure component as part of the Star Rating calculation (**Table 1A**). The results show that while correlations are weaker when the HF EDAC measure is removed from Star Ratings (-0.457 vs. -0.579 for the Readmission Group Score, and -0.349 vs. -0.399 for the Summary Score), most of the relationship between the domains is retained (**Table 1A, Figure 1**).
- We also examined correlations between the HF EDAC measure score and the Star Ratings summary score after removing the entire Readmission group score. We expected a weaker correlation in this case, as we are examining the relationship with measures outside of the Readmission Group, such as those in the domains of Patient Safety, Patient Experience, and Mortality, that are conceptually related to the EDAC outcome (post-discharge hospital visits) but which reflect several different domains of quality. Our results show that there is a weaker, but significant, association between EDAC and Star Ratings (**Table 1B**) even after removing the entire Readmission Group score (-0.199).

Table 1A: Relationship between HF EDAC and Star Ratings

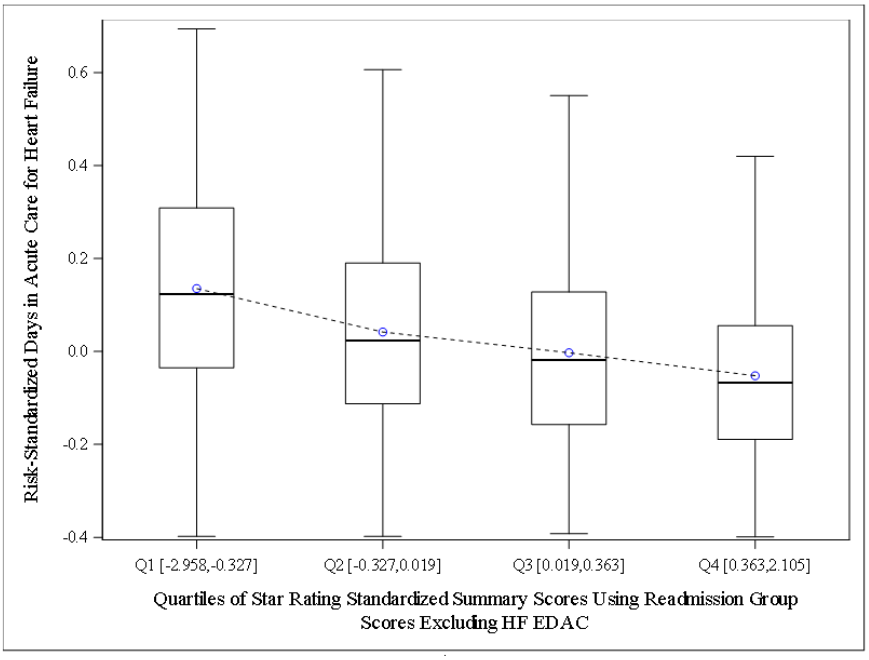
Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	*p* value
Star Rating Standardized Readmission Group Scores	4,264	-0.579	<.0001

Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	*p* value
Star Rating Standardized Readmission Group Scores **Excluding HF EDAC**	4,264	-0.457	<.0001
Star Rating Standardized Summary Scores	4,380	-0.399	<.0001
Star Rating Standardized Summary Scores Using Readmission Group Scores **Excluding HF EDAC**	4,380	-0.349	<.0001

Table 1B: Relationship Between Pneumonia EDAC and Star Ratings without the Readmission Domain

Measure Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	*p* value
Star Rating Standardized Summary Scores **excluding entire Readmission Group Score**	4,420	-0.199	<.0001

Figure 1: Association between quintiles of the Star Ratings Standardized Readmission Group Scores, excluding EDAC, and HF EDAC measure scores



- **Relationship with HCAHPS**
- We expected a relatively weaker correlation with HCAHPS because HCAHPS is a broad measure that captures all patients over age 18 with all clinical conditions, and the EDAC measure captures only patients 65 and older with a specific clinical condition. For HCAHPs, we examined the relationship with the linear mean score or performance score for specific domains, and therefore we would expect a negative correlation (lower EDAC scores for hospitals performing better on HCAHPS domains).
- As hypothesized, while some of the relationships with HCAHPS were less strong than the relationship with the Star Ratings Readmission Group score (without EDAC), we found weak-to-moderate, significant correlations in the expected direction between the HF EDAC measure and components of the HCAHPS survey, including Care Transition performance rates, Doctor, and Nurse Communication linear mean scores, and Discharge Information linear mean scores (Table 2). The analysis presented here used HCAHPS data from calendar year 2018. See Appendix A for details of the HCAPHS components, including the linear mean score.

Table 2: Relationship between HF EDAC and HCAHPS

Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlations with EDAC	*p* value
HCAHPS Care Transition Performance Rates	2,724	-0.237	<.0001

Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlations with EDAC	*p* value
HCAHPS Nurse communication linear mean score	3,328	-0.304	<.0001
HCAHPS Doctor communication linear mean score	3,328	-0.298	<.0001
HCAHPS Staff responsiveness linear mean score	3,328	-0.350	<.0001
HCAHPS Discharge information linear mean score	3,328	-0.314	<.0001

- **Issue 2:** A Panel member observed that the model over-estimates risk in the highest decile and under-estimates risk in the lowest decile.
    - **Developer Response 2:** We also observed the relatively bigger discrepancy in the highest decile, which was associated with the assumption of variance for negative binomial-2 (NB-2) parametrization built in the SAS hierarchical logistic regression procedure (PROC GLIMMIX). Specifically, the variance=mu*(1 + phi*mu), where mu is the mean and bigger for higher deciles and variance increases with the increase of mu. We tested a different parametrization, NB-1, variance = mu*(1+phi'/mu), which could effectively resolve the big offset in the highest decile; however, the offset got bigger for all the rest deciles, especially for the lowest decile. Furthermore, using NB-1 encountered convergence issues. We also tested different distributions (e.g., Poisson, zero-inflated Poisson, zero-inflated NB) and we explored whether including additional factors (e.g., number of comorbidities, Charlson comorbidity index, interactions between risk factors) for risk adjustment would help, but we did not see any significant improvement to this issue. In sum, we have not found a simple solution yet, though we are still exploring different methods and strategies.

- **Issue 3:** Panel members were concerned that the c-statistic of the model was low. There was also concern that the R^2 value was low.
    - **Developer Response 3:** While the c-statistics appear low in comparison to mortality measures, c-statistics in this range are typical for measures like EDAC and readmission. In these cases, quality is driven less by patient characteristics than other characteristics, such as the quality of the facility. CMS's 30-day readmission measures were recently recommended for re-endorsement by the Admissions & Readmissions Standing Committee.
    - We note that medical records-based readmission models perform similarly to the claims-based models: for example, a heart failure readmission model based on medical records had an AUC of 0.58.[1]

- As noted in our submission, the deviance R^2 value was not intended to be used to assess the performance of the risk model in comparison to the risk models for other measures, but rather as way to present how much variance was accounted for by patients' clinical risk factors that hospitals could not control. A low R^2 value in this case indicates that the model only adjusts for a small proportion of variance associated with patients' clinical risk unrelated to hospital performance.
- Note that re-selecting risk variables is currently planned; the original timeline for risk variable re-selection was delayed due to the impact of COVID on work by the developer.

- **Issue 4:** One Panel member thought that model fitting was not reported for a validation data set.
  - **Developer Response 4:** As stated in the text in section 2b3.7, during original measure development we utilized a development and validation data set in our testing, and we present the overfitting indices in the testing attachment. Since this is endorsement maintenance, we only provide model performance statistics on the actual reported cohort. As noted in the testing attachment, 2b3.5, however, we do extensive annual testing to ensure the validity of the model for use with updated data, including updating ICD-10 codes, and evaluating the stability of the risk-adjustment models 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.
- **Issue 5:** One Panel member noted that while the measure scores were shown by the developer to be statistically significantly different, they were unsure if the measure detects "meaningful" differences.
  - **Developer Response 5:** This measure is important to hospitals to be able to understand details about their use of emergency, observation, and readmissions and compare their results to other hospitals. Hospitals can then use the data to implement changes in processes to address areas of concern.
- **Issue 6:** One panel member questioned the rationale behind CMS deciding not to adjust the measure for social risk factors and thought the developer should have presented a reclassification analysis.
  - **Developer Response 6:** We have performed a reclassification analysis on the heart failure readmission measure, rather than the EDAC measure, due the complexity of the EDAC model and the time needed to run the models for this analysis. Our results show that a small number of hospitals are reclassified from one performance category to another. For example, for the dual eligibility variable, only 18 of 3,713 hospitals, or 0.48%, shifted by one performance category (between "better than," "no different than", or "no worse than" the national average) when comparing measures scores with and without adjustment for the dual eligibility variable.
- **Issue 7:** One Panel member asked why the measure did not excluding patients *readmitted* for VAD implantation.
  - **Developer Response 7:** The readmission/admission for an VAD implantation, if not coded with an acute diagnosis, would not be counted as an outcome because it would be considered a planned readmission as per CMS's planned readmission algorithm (PRA)

which is used in this measure. Only unplanned readmissions are counted in the outcome.

- **Issue 8:** A Panel member asked why Medicare Advantage patients and Medicare patients aged <65 are not included the measure.
    - **Developer Response 8:** Medicare Advantage admissions are not included. CMS is currently reviewing the reliability and validity of Medicare Advantage data for use in its measures. Medicare FFS beneficiaries <65 differ from a demographic and clinical perspective from beneficiaries over 65 so there would be a concern regarding adequate risk adjustment.
- **Issue 9:** A Panel member asked if the model was designed to distinguish between a patient with multiple readmissions and multiple patients with single readmission.
    - **Developer Response 9:** EDAC measures are hospital-level measures and we empirically determined that index admissions with multiple unplanned readmissions did not significantly impact hospital readmission rates. Specifically, we found that the hospital readmission rates with or without accounting for multiple unplanned readmissions were highly correlated, which reduced our concern that not accounting for this factor would be problematic. In addition, the current measure uses a complicated two-level hurdle model using the MCMC approach to estimate the hospital-level random effects. Adding a within-patient level effect would exponentially increase the model's complexity but may also cause other potential issues because the vast majority of patients have no or a single unplanned readmission. Therefore, due to the trade-offs, our current measures do not include this feature in the model. However, we may revisit the concern periodically to determine whether it is necessary to revise the measures accordingly.

Appendix A: Details of the HCAHPS Survey Domains

The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey is an NQF-endorsed, publicly reported survey of patients' perspectives of hospital care. Discharged patients (all patients over 18; not limited to Medicare beneficiaries) are asked 29 questions about their hospital experiences, including questions related to communication with doctors and nurses, responsiveness of staff, communication about medications, and discharge information. Scores used for this analysis are based on response categories (for example "No" or "Yes"), the score rate (which is based on the two top box responses), or the linear mean score, which is an average of item-level responses for each question.

Nurse and doctor communication linear mean scores are based on responses to three questions: (1) During this hospital stay, how often did nurses [or doctors for the doctor communication composite] treat you with courtesy and respect? (2) During this hospital stay, how often did nurses [or doctors] listen carefully to you? (3) During this hospital stay, how often did nurses [or doctors] explain things in a way you could understand?

The Discharge Information linear mean score is a composite calculated from responses to two discharge-related questions: (1) During this hospital stay, did doctors, nurses or other hospital staff talk with you about whether you would have the help you needed when you left the hospital? (2) During this hospital stay, did you get information in writing about what symptoms or health problems to look out for after you left the hospital?

The Care Transition Measure (CTM3) is a three-question measure, administered within HCAHPS, that asks the following questions: (1) During this hospital stay, staff took my preferences and those of my family or caregiver into account in deciding what my healthcare needs would be when I left. (2) When I

left the hospital, I had a good understanding of the things I was responsible for in managing my health. (3) When I left the hospital, I clearly understood the purpose for taking each of my medications.

For more information on HCAHPS, see: https://hcahpsonline.org/en/#AboutTheSurvey. For more information on HCAHPS scoring, see: https://hcahpsonline.org/globalassets/hcahps/star-ratings/tech-notes/2017-10_star-ratings_tech-notes.pdf. For more information on the CTM measure, see: https://caretransitions.org/wp-content/uploads/2015/08/CTM3Specs0807.pdf

References

1. Keenan, P. S., Normand, S. L., Lin, Z., Drye, E. E., Bhat, K. R., Ross, J. S., Schuur, J. D., Stauffer, B. D., Bernheim, S. M., Epstein, A. J., Wang, Y., Herrin, J., Chen, J., Federer, J. J., Mattera, J. A., Wang, Y., & Krumholz, H. M. (2008). An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation. Cardiovascular quality and outcomes, 1(1), 29–37.

Measure Number: 2881

Measure Title: Excess days in acute care (EDAC) after hospitalization for acute myocardial infarction (AMI)

Measure Developer: Yale/CORE; Steward: CMS

We thank Scientific Methods Panel members for their thoughtful comments and for the time they took to review the measure specifications and testing results. Below we provide responses to questions posed by the Panel members within their Preliminary Assessment.

Reliability

- **Issue 1:** One Panel member asked about the impact of excluding AMI admissions within 30 days of a prior AMI admission.
    - **Developer Response 1:** Admissions for a condition within 30 days of discharge from an index admission for that same condition are excluded as index admissions. Thus, no hospitalization will be considered as both a readmission and an index admission within the same measure.
    - For AMI, this excludes only 7,173 of 482,163 admissions, or about 1.4%. This exclusion is aligned with exclusions in the respective readmission measures
- **Issue 2:** One Panel member expressed concern over the Planned Readmission Algorithm (PRA) Version 4.0 2020 not clearly eliminating a planned admission for PCI after AMI hospitalization. The Panel member was concerned that if a patient were admitted with any diagnosis such as ICD 10 124.9: Acute ischemic heart disease unspecified, etc. the algorithm would consider diagnosis acute even if the PCI were planned.
    - **Developer Response 2:** The planned readmission algorithm was developed as a means of using administrative data to identify procedures after discharge that were part of the patient's continuation of care. During the development of the planned readmission algorithm, we looked at principal diagnosis codes associated with PCI 30-days post discharge. More than 70% of the time they were associated with principal diagnosis codes falling into the AHRQ CCS group of "Coronary atherosclerosis and other heart

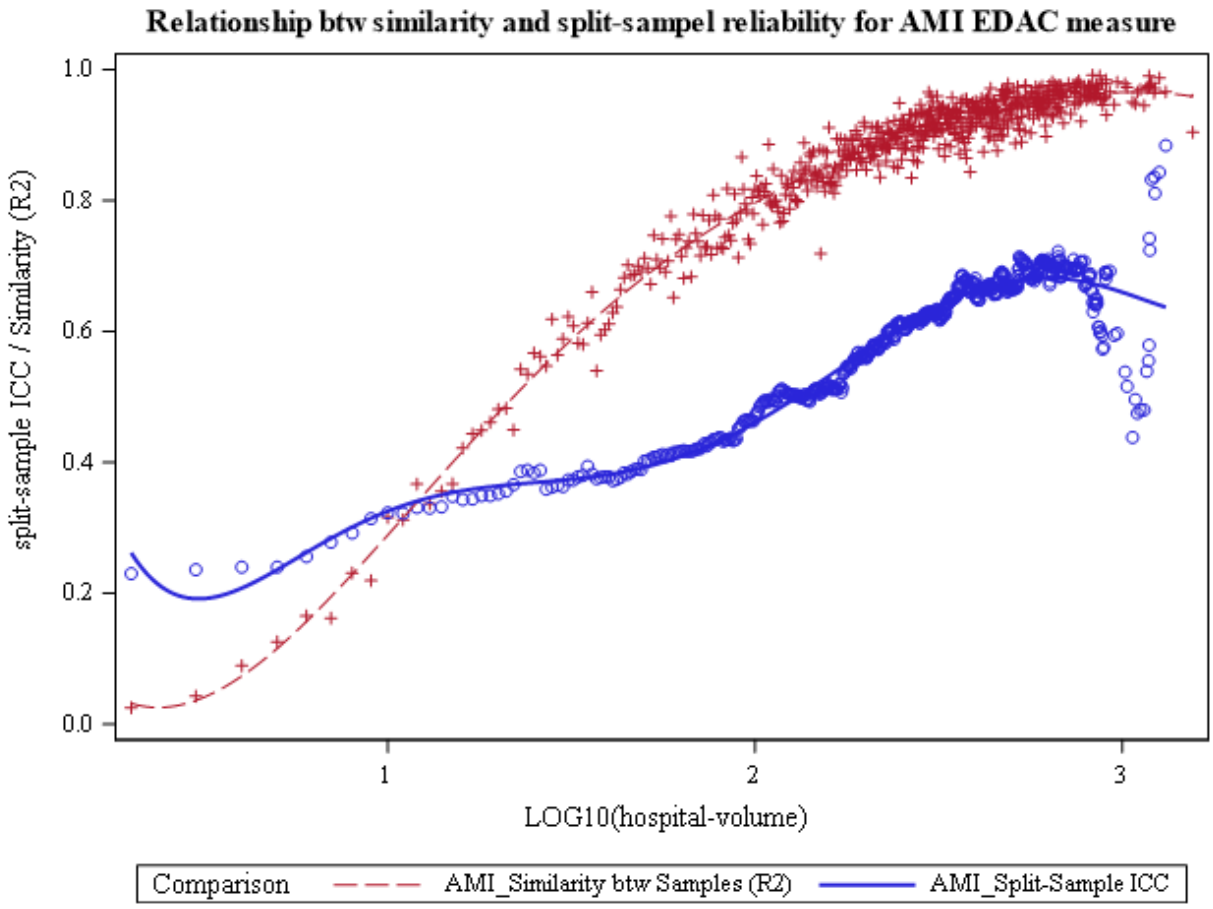
disease" and were determined to be planned. We acknowledge that it is possible that a hospital would use a principal diagnosis code of acute AMI if readmitting a patient for a staged PCI, but our experience is that it is not frequently the case. CMS is aware that the PRA is not 100% accurate, however, our analyses and a chart-review validation study has shown it to be correct in most circumstances. Please note that the PRA is maintained and updated annually based on stakeholder feedback and clinical review.

- **Issue 3:** Some Panel members were concerned that the reliability was lower for this measure than the other EDAC measures. One Panel member evaluated the split-sample reliability results for the AMI EDAC measure against a threshold of 0.7.
    - **Developer Response 3:** Any threshold for reliability would, at a minimum, need to take into account the type of reliability testing that is performed. In our submission, we provided split-sample reliability; signal-to-noise reliability cannot be calculated using a Hurdle model, as outlined in our submission. Split-sample reliability represents the lower bound of reliability [1] and these results are consistent with other reliability measurements in a similar setting [2,3,4] and with the related condition-specific 30-day readmission measures which were recently recommended for NQF re-endorsement.
    - Split-sample reliability is the extent to which repeated measurements of the same hospital give similar results. Accordingly, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produce similar measures of hospital performance. Hospital performance is measured once using a random subset of patients, and then measured again using a second random subset exclusive of the first, and the agreement of the two resulting performance measures is compared across hospitals.[5] We show below, empirically (see **Figure 1**), that the lower reliability for this measure can in part be attributed to the lower similarity between the two split samples when including providers with lower volume. To help illustrate this, in **Figure 1** we show the relationship between hospital volume and the similarity between the two randomly-split samples, and the association of split-sample reliability with the between-sample similarity. In this analysis, we calculate the correlations of risk factors' frequencies (%) between two randomly split samples for each hospital and used the squared-correlations as the indicator of between-sample similarity for each hospital volume. Among hospitals with the same volume, we used the mean squared-correlation as the between-sample similarity for that hospital volume. Next, we calculated the split-sample reliability of each cutoff for hospital volume starting from 2 to the maximum hospital volume (max hospital volume = 1,556 for AMI EDAC). Finally, we plotted both the between-sample similarity and split-sample reliability against hospital volume to visualize the concept that lower reliability is due to lower between-sample similarity when small hospitals were included.
    - A cutoff of 80% (a squared-correlation of 0.80) for between-sample similarity will include hospitals with about 100 or more admissions, which represents split-sample reliabilities of 0.47 or above; while a cutoff of 90% (a squared-correlation of 0.90) for between-sample similarity will include hospitals with about 200 or more admissions,

which represents split-sample reliabilities of 0.58 or above. In sum, as long as the between-sample similarity is sufficient (90%) and fulfills the assumption of split-sample reliability, the split-sample reliability score is within the range of reliability of similar measures (AMI readmission, CABG readmission) recently recommended for NQF re-endorsement.

Figure 1: ICC as function of volume
The assumption of split sample reliability is that the two samples are equivalent or at least as similar as possible. However, empirically and theoretically, the smaller a group is, the less possible it is to split it into two similar samples. The essential challenge for a small-cohort measure is that it is more difficult to get two equivalent or similar samples, which will directly degrade the split-sample reliability.

Here, we visualize the relationship between hospital volumes and the similarity between two randomly split samples along with the association of the similarity between samples with the split-sample reliability. In this way, we hope reviewers can visualize the underlying cause of the lower reliability, which is naturally a feature of the target population and therefore not completely attributable to the measure.



Validity

- **ssue 1:** Panel members expressed concern regarding the validity testing because the approach examined correlation of the EDAC measures with other measures that include the same readmission events, without correction for the overlap. Another Panel member indicated that process-outcome correlations or pre-post analyses of intervention effects are strongly preferred.
    - **Developer Response 1:** As noted in our testing attachment, developers often do not have access to the type of data that would ideally be used for the purposes of empiric validity testing, such as patient-level data on process measures that are related to the outcome.

- We have, however, included updated testing results that implement the methodologic approach suggested by the Scientific Methods Panel with respect to Star Ratings (removing the comparator measure from the Star Rating Readmission Group score before analyzing the correlation). In addition, because conceptually there could be a relationship between EDAC and the other domains of Star Ratings (which include measures related to care coordination in the domain of Patient Experience, hospital-acquired infections in the Safety of Care domain, in addition to measures related to timeliness and effective care in the Emergency Department, as well as mortality) we also examined the association between the EDAC measure and Star Ratings after removing the entire Readmission Group score. Finally, we separately examined the relationship between AMI EDAC and components of the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, a validated and NQF-endorsed patient-reported quality measure that reflects processes of care that relate, conceptually, to post-discharge hospital visits.
- Our results show that, while weaker, there remains a correlation in the expected direction between the EDAC measure and Star Ratings even after removing the overlapping measure (Table 1A). We also find an association, albeit weaker, between the EDAC measure and Star Ratings after removing the entire Readmission Group from Star Ratings (Table 1B), suggesting that the EDAC measures share a quality signal with the remaining Star Ratings domains, which include measures related to patient safety, timeliness and effective care, patient experience, and mortality.
- Consistent with the results seen for Star Ratings, our results show an association in the hypothesized direction with components of HCAHPS (Table 2). Taken together, the results provide additional support for the validity of the AMI EDAC measure. The results are discussed in more detail below.
- **Relationship with Star Ratings**
- Briefly, using the **recently updated Star Ratings methodology** that no longer uses the latent variable model, but rather averages the performance of measures in the measure group, we compared hospitals' performances on AMI EDAC with performance on Star Ratings summary scores and Readmission Group scores, with and without the AMI EDAC measure component as part of the Star Rating calculation (**Table 1A**). The results show that while correlations are weaker when the AMI EDAC measure is removed from Star Ratings (-0.313 vs. -0.380 for the Readmission Group Score, and -0.221 vs. -0.247 for the Summary Score), most of the association between the domains is retained (**Table 1A**, **Figure 1**).
- We also examined correlations between the AMI EDAC measure score and the Star Ratings summary score after removing the entire Readmission group score. We expected a weaker correlation in this case, as we are examining the relationship with measures outside of the Readmission Group, such as those in the domains of Patient Safety, Patient Experience, and Mortality, that are conceptually related to the EDAC outcome (post-discharge hospital visits) **but which reflect several different domains of quality**. Our results show that there is a weaker, but significant, association between

EDAC and Star Ratings (Table 1B) even after removing the entire Readmission Group score (-0.120). Note that the strength of the correlation is not the focus of this analysis; rather we are interested in knowing if the calculated correlation coefficient aligns with the expected conceptual relationship between the two measures.

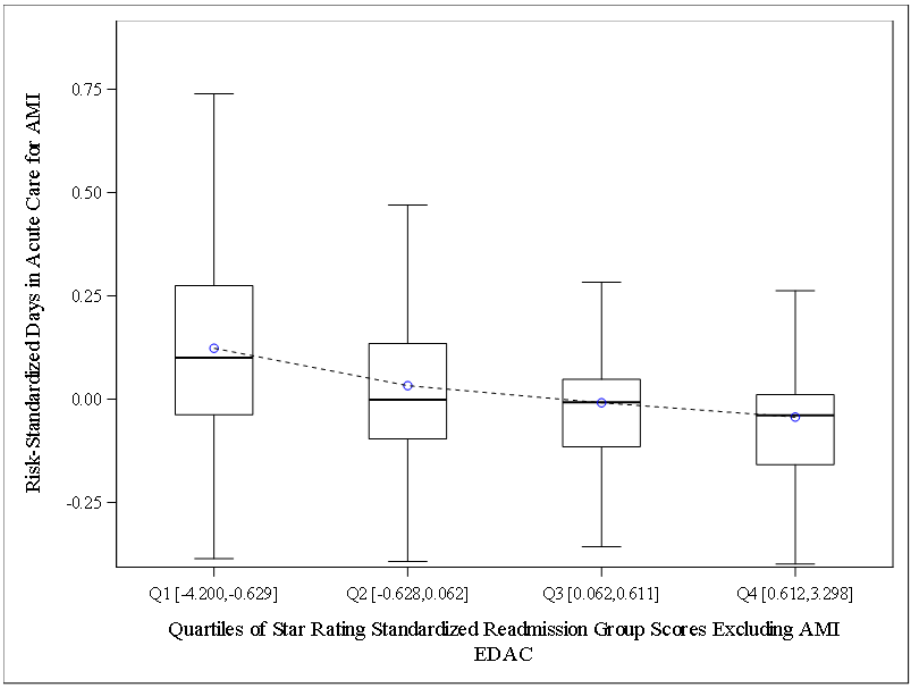
Table 1A: Relationship between AMI EDAC and Star Ratings

Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	p value
Star Rating Standardized Readmission Group Scores	3,840	-0.380	<.0001
Star Rating Standardized Readmission Group Scores **Excluding AMI EDAC**	3,840	-0.313	<.0001
Star Rating Standardized Summary Scores	3,877	-0.247	<.0001
Star Rating Standardized Summary Scores Using Readmission Group Scores **Excluding AMI EDAC**	3,877	-0.221	<.0001

Table 1B: Relationship Between Pneumonia EDAC and Star Ratings without the Readmission Domain

Measure Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	p value
Star Rating Standardized Summary Scores **excluding entire Readmission Group Score**	3,877	-0.120	<.0001

Figure 1: Association between quintiles of the Star Ratings Standardized Readmission Group Scores, excluding EDAC, and AMI EDAC measure scores



- **Relationship with HCAHPS**
- We expected a relatively weaker correlation with HCAHPS because HCAHPS is a broad measure that captures all patients over age 18 with all clinical conditions, and the EDAC measure captures only patients 65 and older with a specific clinical condition. For HCAHPs, we examined the relationship with the linear mean score or performance score for specific domains, and therefore we would expect a negative correlation (lower EDAC scores for hospitals performing better on HCAHPS domains).
- As hypothesized, while some of the relationships with HCAHPS were less strong than the relationship with the Star Ratings Readmission Group score (without EDAC), we found moderate significant correlations in the expected direction between the AMI EDAC measure and components of the HCAHPS survey, including Care Transition performance rates, Doctor, and Nurse Communication linear mean scores, and Discharge Information linear mean scores (Table 2). The analysis presented here used HCAHPS data from calendar year 2018. See Appendix A for details of the HCAPHS components, including the linear mean score.

Table 2: Relationship between AMI EDAC and HCAHPS

Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	*p* value
HCAHPS Care Transition Performance Rates	2,686	-0.189	<.0001

Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	*p* value
HCAHPS Nurse communication linear mean score	3,229	-0.213	<.0001
HCAHPS Doctor communication linear mean score	3,229	-0.195	<.0001
HCAHPS Staff responsiveness linear mean score	3,229	-0.208	<.0001
HCAHPS Discharge information linear mean score	3,229	-0.232	<.0001

- **Issue 2:** A Panel member observed that the model over-estimates risk in the highest decile and under-estimates risk in the lowest decile.
    - **Developer Response 2:** We also observed the relatively bigger discrepancy in the highest decile, which was associated with the assumption of variance for negative binomial-2 (NB-2) parametrization built in the SAS hierarchical logistic regression procedure (PROC GLIMMIX). Specifically, the variance=mu*(1 + phi*mu), where mu is the mean and bigger for higher deciles and variance is increases with the increase of mu. We tested a different parametrization, NB-1, variance = mu*(1+phi'/mu), which could effectively resolve the big offset in the highest decile; however, the offset got bigger for all the rest deciles, especially for the lowest decile. Furthermore, using NB-1 encountered convergence issues. We also tested different distributions (e.g., Poisson, zero-inflated Poisson, zero-inflated NB) and we explored whether including additional factors (e.g., number of comorbidities, Charlson comorbidity index, interactions between risk factors) for risk adjustment would help, but we did not see any significant improvement to this issue. In sum, we have not found a simple solution yet, though we are still exploring different methods and strategies.
- **Issue 3:** Panel members were concerned that the c-statistic of the model was low.
    - **Developer Response 3:** While the c-statistics appear low in comparison to mortality measures, c-statistics in this range are typical for measures like EDAC and readmission. In these cases, quality is driven less by patient characteristics than other characteristics, such as the quality of the facility. CMS's 30-day readmission measures were recently recommended for re-endorsement by the Admissions & Readmissions Standing Committee.
    - We note that medical-records based readmission models perform similarly to the claims-based models: for example, an AMI readmission model based on medical records had an AUC of 0.58.[6]

      As noted in our submission, the deviance R^2 value was not intended to be used to assess the performance of the risk model in comparison to the risk models for other measures, but rather as way to present how much variance was accounted for by patients' clinical

risk factors that hospitals could not control. A low R^2 value in this case indicates that the model only adjusts for a small proportion of variance associated with patients' clinical risk unrelated to hospital performance.
Note that re-selecting risk variables is currently planned; the original timeline for risk-variable re-selection was delayed due to the impact of COVID on work by the developer.

- **Issue 4:** One Panel member thought that model fitting was not reported for a validation data set.
    - **Developer Response 4:** As stated in the text in section 2b3.7, during original measure development we utilized a development and validation data set in our testing, and we present the overfitting indices in the testing attachment. Since this is endorsement maintenance, we only provide model performance statistics on the actual reported cohort. As noted in the testing attachment, 2b3.5, however, we do extensive annual testing ensure the validity of the model for use with updated data, including updating ICD-10 codes, and evaluating the stability of the risk-adjustment models 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.
- **Issue 5:** One panel member noted that while the measure scores were shown by the developer to be statistically significantly different, they were unsure if the measure detects "meaningful" differences.
    - **Developer Response 5:** This measure is important to hospitals to be able to understand details about their use of emergency, observation, and readmissions and compare their results to other hospitals. Hospitals can then use the data to implement changes in processes to address areas of concern.
- **Issue 6:** One Panel member questioned the rationale behind CMS deciding not to adjust the measure for social risk factors and thought the developer should have presented a reclassification analysis.
    - **Developer Response 6:** We have performed a reclassification analysis on the heart failure readmission measure, rather than the EDAC measure, due the complexity of the EDAC model and the time needed to run the models for this analysis. Our results show that a small number of hospitals are reclassified from one performance category to another. For example, for the dual eligibility variable, only 18 of 3,713 hospitals, or 0.48%, shifted by one performance category (between "better than," "no different than", or "no worse than" the national average) when comparing measures scores with and without adjustment for the dual eligibility variable.

Appendix A: Details of the HCAHPS Survey Domains

The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey is an NQF-endorsed, publicly reported survey of patients' perspectives of hospital care. Discharged patients (all patients over 18; not limited to Medicare beneficiaries) are asked 29 questions about their hospital experiences, including questions related to communication with doctors and nurses, responsiveness of staff, communication about medications, and discharge information. Scores used for this analysis are based on response categories (for example "No" or "Yes"), the score rate (which is based on the two top box responses), or the linear mean score, which is an average of item-level responses for each question.

Nurse and doctor communication linear mean scores are based on responses to three questions: (1) During this hospital stay, how often did nurses [or doctors for the doctor communication composite] treat you with courtesy and respect? (2) During this hospital stay, how often did nurses [or doctors] listen carefully to you? (3) During this hospital stay, how often did nurses [or doctors] explain things in a way you could understand?

The Discharge Information linear mean score is a composite calculated from responses to two discharge-related questions: (1) During this hospital stay, did doctors, nurses or other hospital staff talk with you about whether you would have the help you needed when you left the hospital? (2) During this hospital stay, did you get information in writing about what symptoms or health problems to look out for after you left the hospital?

The Care Transition Measure (CTM3) is a three-question measure, administered within HCAHPS, that asks the following questions: (1) During this hospital stay, staff took my preferences and those of my family or caregiver into account in deciding what my healthcare needs would be when I left. (2) When I left the hospital, I had a good understanding of the things I was responsible for in managing my health. (3) When I left the hospital, I clearly understood the purpose for taking each of my medications.

For more information on HCAHPS, see: https://hcahpsonline.org/en/#AboutTheSurvey. For more information on HCAHPS scoring, see: https://hcahpsonline.org/globalassets/hcahps/star-ratings/tech-notes/2017-10_star-ratings_tech-notes.pdf. For more information on the CTM measure, see: https://caretransitions.org/wp-content/uploads/2015/08/CTM3Specs0807.pdf

References

1. Adams J. The reliability of provider profiling, a tutorial. RAND Corporation, 2009. Available at: https://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR653.pdf. Accessed on March 9, 2021.
2. 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.
3. 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.
4. 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.
5. Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. Statistics in medicine, 21(22), 3431–3446.D
6. Krumholz, H. M., Lin, Z., Drye, E. E., Desai, M. M., Han, L. F., Rapp, M. T., Mattera, J. A., & Normand, S. L. (2011). An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Circulation. Cardiovascular quality and outcomes, 4(2), 243–252.

Measure Number: 2882

Measure Title: Excess days in acute care (EDAC) after hospitalization for pneumonia

Measure Developer: Yale/CORE; Steward: CMS

We thank Scientific Methods Panel members for their thoughtful comments and for the time they took to review the measure specifications and testing results. Below we provide responses to questions posed by the Panel members within their Preliminary Assessment.

Validity

- **Issue 1:** Panel members expressed concern regarding the validity testing because the approach examined correlated of the EDAC measures with other measures that include the same readmission events, without correction for the overlap. Another Panel member indicated that process-outcome correlations or pre-post analyses of intervention effects are strongly preferred.
    - **Developer Response 1:** As noted in our testing attachment, developers often do not have access to the type of data that would ideally be used for the purposes of empiric validity testing, such as patient-level data on process measures that are related to the outcome.
    - We have, however, included updated testing results that implement the methodologic approach suggested by the Scientific Methods Panel with respect to Star Ratings (removing the comparator measure from the Star Rating Readmission Group score before analyzing the correlation). In addition, because conceptually there could be a relationship between EDAC and the other domains of Star Ratings (which include measures related to care coordination in the domain of Patient Experience, hospital-acquired infections in the Safety of Care domain, in addition to measures related to timeliness and effective care in the Emergency Department, as well as mortality) we also examined the association between the EDAC measure and Star Ratings after removing the entire Readmission Group score. Finally, we separately examined the relationship between pneumonia EDAC and components of the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, a validated and NQF-endorsed patient-reported quality measure that reflects processes of care that relate, conceptually, to post-discharge hospital visits.
    - Our results show that there remains a moderate correlation in the expected direction between the EDAC measure and Star Ratings even after removing the overlapping measure (Table 1A). We also find an association, albeit weaker, between the EDAC measure and Star Ratings after removing the entire Readmission Group from Star Ratings (Table 1B), suggesting that the EDAC measures share a quality signal with the remaining Star Ratings domains, which include measures related to patient safety, timeliness and effective care, patient experience, and mortality.
    - Consistent with the results seen for Star Ratings, our results show an association in the hypothesized direction with components of HCAHPS (Table 2). Taken together, the results provide additional support for the validity of the pneumonia EDAC measure. The results are discussed in more detail below.
    - **Relationship with Star Ratings**
    - Briefly, using the recently **updated Star Ratings methodology** that no longer uses the latent variable model, but rather averages the performance of measures in the measure

group, we compared hospitals' performances on pneumonia EDAC with performance on Star Ratings summary scores and Readmission Group scores, with and without the pneumonia EDAC measure component as part of the Star Rating calculation (**Table 1A**). The results show that while correlations are weaker when the pneumonia EDAC measure is removed from Star Ratings (-0.451 vs. -0.618 for the Readmission Group score, and -0.363 vs. -0.437 for the Summary Score), most of the association between the measures is retained (**Table 1A, Figure 1**).

- We also examined correlations between the pneumonia EDAC measure score and the Star Ratings summary score after removing the entire Readmission group score. We expected a weaker correlation in this case, as we are examining the relationship with measures outside of the Readmission Group, such as those in the domains of Patient Safety, Patient Experience, and Mortality, that are conceptually related to the EDAC outcome (post-discharge hospital visits) but which reflect several different domains of quality. Our results show that there is a weaker, but significant, association between EDAC and Star Ratings (**Table 1B**) even after removing the entire Readmission Group score (-0.221).

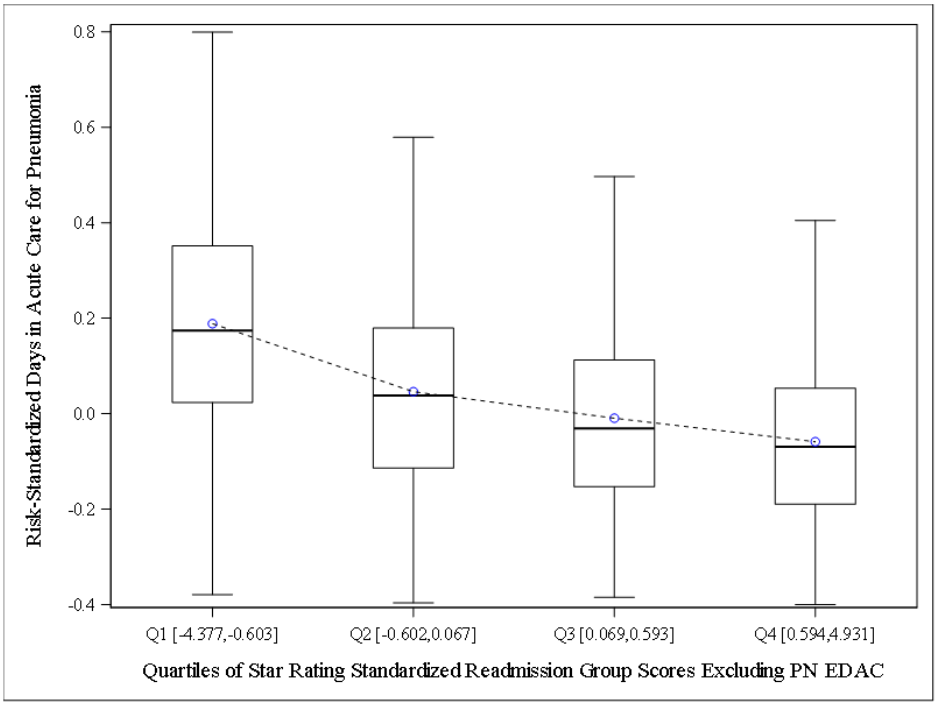
Table 1A: Relationship Between Pneumonia EDAC and Star Ratings with Readmission Component

Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	*p* value
Star Rating Standardized Readmission Group Scores	4,279	-0.618	<.0001
Star Rating Standardized Readmission Group Scores **excluding pneumonia EDAC**	4,277	-0.451	<.0001
Star Rating Standardized Summary Scores	4,420	-0.437	<.0001
Star Rating Standardized Summary Scores Using Readmission Group Scores **excluding pneumonia EDAC**	4,420	-0.363	<.0001

Table 1B: Relationship Between Pneumonia EDAC and Star Ratings without the Readmission Domain

Measure Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	*p* value
Star Rating Standardized Summary Scores **excluding entire Readmission Group Score**	4,420	-0.221	<.0001

Figure 1: Association between quintiles of the Star Ratings Standardized Readmission Group Scores, excluding EDAC, and Pneumonia EDAC measure scores



Relationship with HCAHPS

We expected a relatively weaker correlation with HCAHPS because HCAHPS is a broad measure that captures all patients over age 18 with all clinical conditions, and the EDAC measure captures only patients 65 and older with a specific clinical condition. For HCAHPs, we examined the relationship with the linear mean score or performance score for specific domains, and therefore we would expect a negative correlation (lower EDAC scores for hospitals performing better on HCAHPS domains).

As hypothesized, while some of the relationships with HCAHPS were less strong than the relationship with the Star Ratings Readmission Group score (without EDAC), we found moderate significant correlations in the expected direction between the pneumonia EDAC measure and components of the HCAHPS survey, including Care Transition performance rates, Doctor, and Nurse Communication linear mean scores, and Discharge Information linear mean scores (Table 2). The analysis presented here used

HCAHPS data from calendar year 2018. See **Appendix A** for details of the HCAPHS components, including the linear mean score.

Table 2: Relationship between pneumonia EDAC and HCAHPS

Measures Used for Validity Testing	Number of Hospitals	Pearson's Correlation with EDAC	*p* value
HCAHPS Care Transition (CTM3) Performance Rates	2,726	-0.239	<.0001
HCAHPS Nurse Communication linear mean score	3,342	-0.376	<.0001
HCAHPS Doctor Communication linear mean score	3,342	-0.346	<.0001
HCAHPS Discharge information linear mean score	3,342	-0.356	<.0001

- **Issue 2:** A Panel member observed that the model over-estimates risk in the highest decile and under-estimates risk in the lowest decile.
    - **Developer Response 2:** We also observed the relatively bigger discrepancy in the highest decile, which was associated with the assumption of variance for negative binomial-2 (NB-2) parametrization built in the SAS hierarchical logistic regression procedure (PROC GLIMMIX). Specifically, the variance=mu*(1 + phi*mu), where mu is the mean and bigger for higher deciles and variance is increases with the increase of mu. We tested a different parametrization, NB-1, variance = mu*(1+phi'/mu), which could effectively resolve the big offset in the highest decile; however, the offset got bigger for all the rest deciles, especially for the lowest decile. Furthermore, using NB-1 encountered convergence issues. We also tested different distributions (e.g., Poisson, zero-inflated Poisson, zero-inflated NB) and we explored whether including additional factors (e.g., number of comorbidities, Charlson comorbidity index, interactions between risk factors) for risk adjustment would help, but we did not see any significant improvement to this issue. In sum, we have not found a simple solution yet, though we are still exploring different methods and strategies.
- **Issue 3:** Panel members were concerned that the c-statistic of the model was low. There was also concern that the R^2 value was low.
    - **Developer Response 3:** While the c-statistics appear low in comparison to mortality measures, c-statistics in this range are typical for measures like EDAC and readmission. In these cases, quality is driven less by patient characteristics than other characteristics, such as the quality of the facility. CMS's 30-day readmission measures were recently recommended for re-endorsement by the Admissions & Readmissions Standing Committee.

- As noted in our submission, the deviance R^2 value was not intended to be used to assess the performance of the risk model in comparison to the risk models for other measures, but rather as way to present how much variance was accounted for by patients' clinical risk factors that hospitals could not control. A low R^2 value in this case indicates that the model only adjusts for a small proportion of variance associated with patients' clinical risk unrelated to hospital performance.
- Note that re-selecting risk variables is currently planned; the original timeline for risk-variable re-selection was delayed due to the impact of COVID on work by the developer.

- **Issue 4:** One Panel member thought that model fitting was not reported for a validation data set.
  - **Developer Response 4:** As stated in the text in section 2b3.7, during original measure development we utilized a development and validation data set in our testing, and we present the overfitting indices in the testing attachment. Since this is endorsement maintenance, we only provide model performance statistics on the actual reported cohort. As noted in the testing attachment, 2b3.5, however, we do extensive annual testing to ensure the validity of the model for use with updated data, including updating ICD-10 codes, and evaluating the stability of the risk-adjustment models 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.
- **Issue 5:** One Panel member noted that while the measure scores were shown by the developer to be statistically significantly different, they were unsure if the measure detects "meaningful" differences.
  - **Developer Response 5:** This measure is important to hospitals to be able to understand details about their use of emergency, observation, and readmissions and compare their results to other hospitals. Hospitals can then use the data to implement changes in processes to address areas of concern.
- **Issue 6:** One Panel member questioned the rationale behind CMS deciding not to adjust the measure for social risk factors and thought the developer should have presented a reclassification analysis.
  - **Developer Response 6:** We have performed a reclassification analysis on the heart failure readmission measure, rather than the EDAC measure, due the complexity of the EDAC model and the time needed to run the models for this analysis. Our results show that a small number of hospitals are reclassified from one performance category to another. For example, for the dual eligibility variable, only 18 of 3,713 hospitals, or 0.48%, shifted by one performance category (between "better than," "no different than", or "no worse than" the national average) when comparing measures scores with and without adjustment for the dual eligibility variable.

Appendix A: Details of the HCAHPS Survey Domains

The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey is an NQF-endorsed, publicly reported survey of patients' perspectives of hospital care. Discharged patients (all patients over 18; not limited to Medicare beneficiaries) are asked 29 questions about their hospital experiences, including questions related to communication with doctors and nurses, responsiveness of staff, communication about medications, and discharge information. Scores used for this analysis are

based on response categories (for example "No" or "Yes"), the score rate (which is based on the two top box responses), or the linear mean score, which is an average of item-level responses for each question.

Nurse and doctor communication linear mean scores are based on responses to three questions: (1) During this hospital stay, how often did nurses [or doctors for the doctor communication composite] treat you with courtesy and respect? (2) During this hospital stay, how often did nurses [or doctors] listen carefully to you? (3) During this hospital stay, how often did nurses [or doctors] explain things in a way you could understand?

The Discharge Information linear mean score is a composite calculated from responses to two discharge-related questions: (1) During this hospital stay, did doctors, nurses or other hospital staff talk with you about whether you would have the help you needed when you left the hospital? (2) During this hospital stay, did you get information in writing about what symptoms or health problems to look out for after you left the hospital?

The Care Transition Measure (CTM3) is a three-question measure, administered within HCAHPS, that asks the following questions: (1) During this hospital stay, staff took my preferences and those of my family or caregiver into account in deciding what my healthcare needs would be when I left. (2) When I left the hospital, I had a good understanding of the things I was responsible for in managing my health. (3) When I left the hospital, I clearly understood the purpose for taking each of my medications.

For more information on HCAHPS, see: https://hcahpsonline.org/en/#AboutTheSurvey. For more information on HCAHPS scoring, see: https://hcahpsonline.org/globalassets/hcahps/star-ratings/tech-notes/2017-10_star-ratings_tech-notes.pdf. For more information on the CTM measure, see: https://caretransitions.org/wp-content/uploads/2015/08/CTM3Specs0807.pdf

Measure Number: 3188

Measure Title: 30-Day Unplanned Readmissions for Cancer Patients

Measure Developer: Alliance of Dedicated Cancer Centers

Reliability

We appreciate the review and insight regarding reliability testing for this measure; however, per recommendation of NQF Methods Panel staff, we are not providing comments regarding reliability as the measure passed on reliability.

Validity

- **Issue 1:** Concern with use of the correlation between 30-Day Unplanned Readmissions for Cancer Patients and the CMS hospital-wide all-cause readmission measure (NQF #1789) for performance level validity testing. Specific concerns included endogeneity/ intrinsic correlation between these measures.
    - **Developer Response 1:** Due to ongoing gaps cancer quality measures, we did not identify NQF-endorsed cancer-specific process or outcome measures suitable for this purpose. However, we determined that CMS' Hospital-Wide All-Cause Readmission Measure (HWR) (NQF #1789) could be used for empirical validity testing. The specific rationale to choose the measure are outlined in Section 2b1.2 of the original submission. It is important to note that NQF #1789 specifically excludes patients admitted for medical treatment of cancer (i.e., patients with a principal diagnosis of cancer are excluded unless undergoing a major surgical procedure) and excludes index admissions to the nation's 11 Prospective Payment System

(PPS)-exempt cancer hospitals. The reviewers may not have been aware of the presence of these exclusions in HWR.

- Per the CMS measure specifications for NQF #1789, admissions for the medical treatment of cancer are excluded because 'these admissions have a very different mortality and readmission profile than the rest of the Medicare population, and outcomes for these admissions do not correlate well with outcomes for other admissions.'
- The codes for this exclusion are below (source: the data dictionary on the NQF site).

Codes Currently on NQF Site for NQF #1789

AHRQ CCS Cancer Discharge Diagnosis Categories Excluded from the Measure

AHRQ Diagnosis CCS (ICD-10)	Description
11	Cancer of head and neck
12	Cancer of esophagus
13	Cancer of stomach
14	Cancer of colon
15	Cancer of rectum and anus
16	Cancer of liver and intrahepatic bile duct
17	Cancer of pancreas
18	Cancer of other GI organs; peritoneum
19	Cancer of bronchus; lung
20	Cancer; other respiratory and intrathoracic
21	Cancer of bone and connective tissue
22	Melanomas of skin
23	Other non-epithelial cancer of skin
24	Cancer of breast
25	Cancer of uterus
26	Cancer of cervix
27	Cancer of ovary
28	Cancer of other female genital organs
29	Cancer of prostate
30	Cancer of testis
31	Cancer of other male genital organs
32	Cancer of bladder
33	Cancer of kidney and renal pelvis
34	Cancer of other urinary organs
35	Cancer of brain and nervous system
36	Cancer of thyroid
37	Hodgkin`s disease
38	Non-Hodgkin`s lymphoma
39	Leukemias
40	Multiple myeloma
41	Cancer; other and unspecified primary
42	Secondary malignancies
43	Malignant neoplasm without specification of site
44	Neoplasms of unspecified nature or uncertain behavior
45	Maintenance chemotherapy; radiotherapy

- **Issue 2:** One panelist expressed concern that there was no new testing of data element validity. A panelist indicated that they would score validity more favorably if the originally submitted

data element validity testing was allowable.

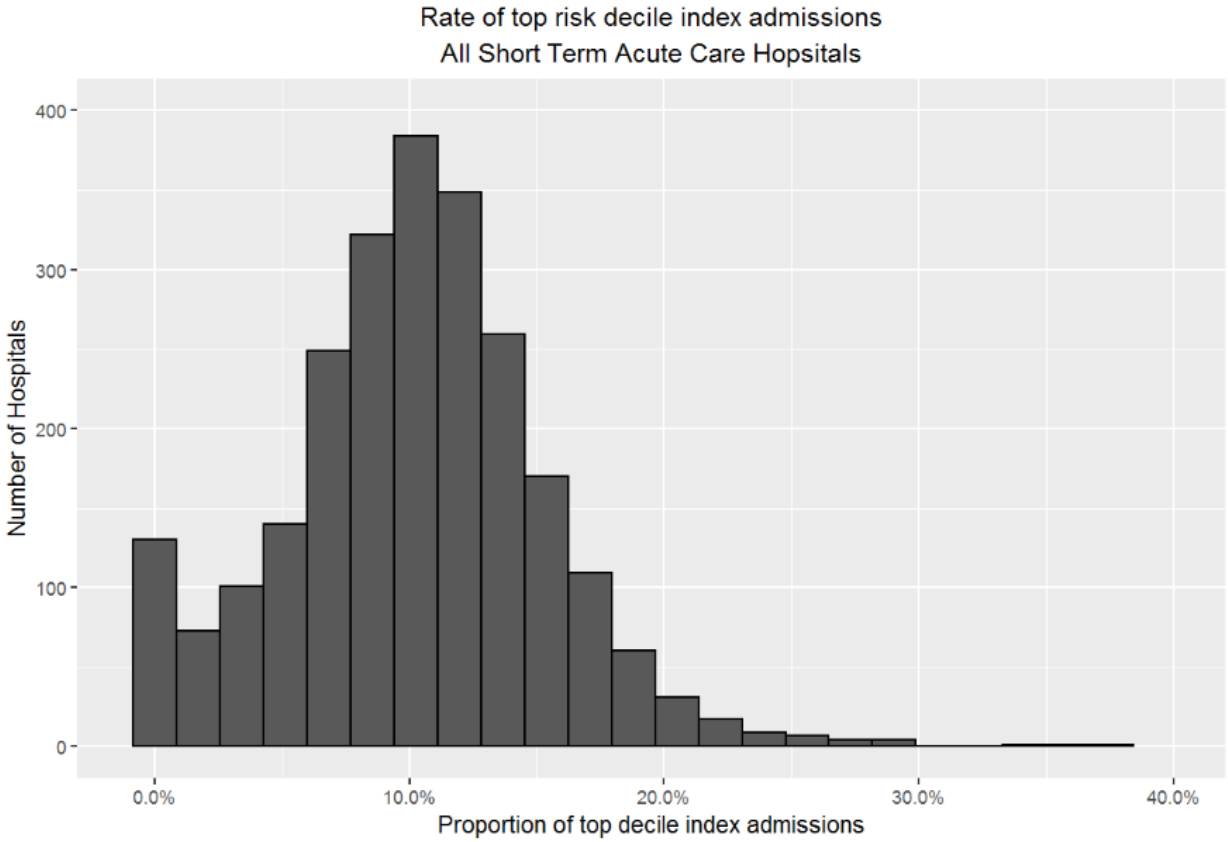
- **Developer Response 2:** The data elements in this measure did not change during the maintenance review and analysis. The NQF document, "Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement - Effective September 2019" indicates that prior testing data may be used if results demonstrated that validity achieved at least a moderate rating. This was the case for measure 3188.

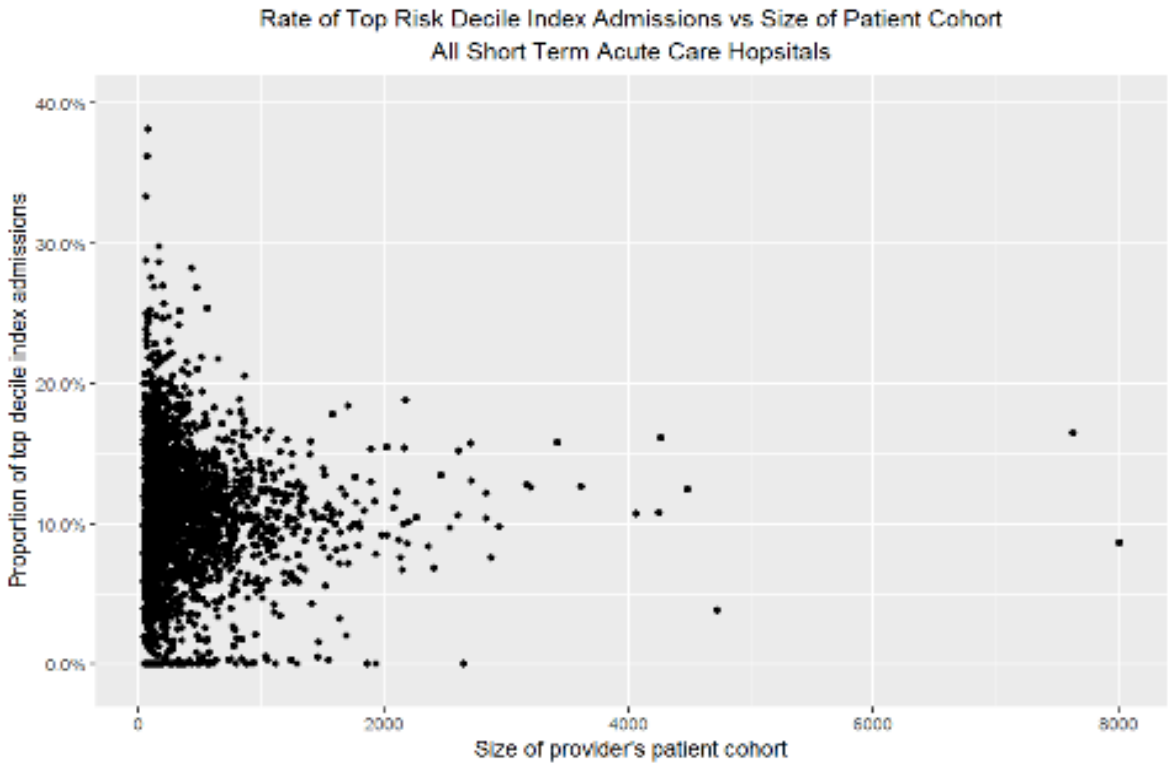
	First Maintenance Evaluation	Subsequent Maintenance Evaluations
Reliability	Measure In Use • Analysis of data from entities whose performance is measured • Reliability of measure scores Measure Not in Use • Expanded testing in terms of scope (number of entities/patients) and/or levels (data elements/measure score)	Could submit prior testing data, if results demonstrated that reliability achieved at least a moderate rating
Validity	Measure in Use • Analysis of data from entities whose performance is measured • Validity of measure score for making accurate conclusions about quality • Updated/expanded analysis of threats to validity Measure Not in Use • Expanded testing in terms of scope (number of entities/patients) and/or levels (data elements/measure score)	Could submit prior testing data, if results demonstrated that validity achieved at least a moderate rating

- **Issue 3:** Concerns about the strength of correlation between this measure and NQF #1789 were expressed.
  - **Developer Response 3:** We found a moderate positive correlation (0.255) between this measure (NQF #3188) and the CMS hospital-wide all-cause readmissions measure (NQF #1789). While we hypothesized a positive correlation, we are not surprised that the correlation was not stronger as there are differences between the general acute care hospital patients, and cancer patients, as was noted in documentation for NQF#1789 (briefly described above). In fact, the unique characteristics of the cancer patient population and cancer-related treatment are compelling reasons supporting the need for this measure.
- **Issue 4:** One reviewer expressed concerns about the risk factors included, and specifically those omitted.
  - **Developer Response 4:** Please see Section 2b3.3a. of the testing form for a detailed discussion of the process used to identify risk factors, evaluate our ability to access data regarding these factors, select variables for testing, and final variables selected. This process follows standard, accepted practices and was considered acceptable during the original NQF review of this measure. This process was repeated for maintenance submission, leveraging teams of clinical, quality and coding experts from the Alliance of Dedicated Cancer Centers (ADCC) Quality Working Group and members of the ADCC Physician Advisory Working

Group.

- **Issue 5:** One panel member expressed that that dual eligibility is the only SDS variable included. Another reviewer questioned the decision to remove race from the risk adjustment model.
  - **Developer Response 5:** Risk adjustment variables included age, gender, and dual eligibility status. Race also was tested, as discussed in section 2b3.3a. As described in 2b3.4b., our exploratory analysis including race detected statistically significant differences among White, Black, and Asian patients, but not among Hispanic, Native American, and Unknown/Other patients. Inclusion of race had limited impact on estimated patient level readmission probabilities or provider level risk-adjusted rates. Comparing risk adjusted models with and without race produced index admission level readmission probabilities with a correlation coefficient of 0.99838 (CI: 0.99837, 0.99839) and provider level risk adjusted rates with a correlation coefficient of 0.99993 (CI: 0.99992, 0.99994).
- **Issue 6:** Visually, the model appeared to under-estimate risk in the highest decile of predicted risk in CY2016. Is the proportion of patients with predicted risk estimates in the highest decile relatively constant across providers?
  - **Developer Response 6:** Yes, the model does appear to somewhat underestimate the risk of readmission for patients in the top risk decile. However, examining the distribution of these patients, we found that patients from this decile appear to be spread randomly across providers, with the distribution of hospitals' proportion of patients from that decile being centered around 10% (sd 5%). Size of cohort and type of provider did not appear to impact a hospitals' proportion of patients. Despite this underestimation, the model appears to provide strong predictive power, given the results of the C statistic, as well as good discrimination between the lowest and highest risk patients as well as between providers, given the rest of the decile plot. While the performance of the model is suboptimal in the top decile, it does not appear to unfairly discriminate across the providers. We appreciate the reviewer's comments on the performance in the top decile and will continue to evaluate the model for ways to improve this.


Rate of top risk decile index admissions
All Short Term Acute Care Hopsitals
Number of Hospitals
400
300
200
100
0
0.0%
10.0%
20.0%
30.0%
40.0%
Proportion of top decile index admissions


Rate of Top Risk Decile Index Admissions vs Size of Patient Cohort
All Short Term Acute Care Hopsitals
Proportion of top decile index admissions
40.0%
30.0%
20.0%
10.0%
0.0%
0
2000
4000
6000
8000
Size of provider's patient cohort

- **Issue 7:** The MIF (S.8) states one of the exclusions as "Patients having missing or incomplete data". In S.9 on the MIF it defines this as "Patient Discharge Status Code indicating "Unknown Value" (0, 40-42) or Organization NPI Number = "". However, the testing form is silent on this & thus doesn't provide any analyses regarding this. In turn, we are not able to evaluate the significance of this missing data.
    - **Developer Response 7:** Section 2b2.2. , Table 2b2.2.A on page 19 of the submitted NQF Testing Attachment reports the number of patients excluded at each exclusion step. This table indicated that the number of patients excluded due to missing or incomplete data was zero. The developers should have clarified in the Testing Attachment that 'patients excluded as having missing or incomplete data' includes Patient Discharge Code Status and Organizational NPI. We confirmed that this table is correct and that no patients were excluded from the measure due missing Patient Discharge Status Code or missing NPI numbers.
- **Issue 8:** One panelist stated that "After several years there should be data to validate that lower readmission correlates with other measures of quality of care in the patient population being studied."
    - **Developer Response 8:** As previously noted, there is a lack of cancer outcome measures which would be appropriate for comparison. Unfortunately, this outcome measure, #3188, has not yet been implemented by CMS in the quality reporting program for which it was adopted, the PPS-Exempt Cancer Hospital Quality Reporting Program. The first release of this data, via a confidential national data reporting (dry run) is anticipated to occur in Summer 2021. It is the developers' hope and expectation that additional outcome measures will be available for this important patient population in the future.

Measure Number: 3612

Measure Title: Risk-Standardized Acute Cardiovascular-Related Hospital Admission Rates for Patients with Heart Failure under the Merit-based Incentive Payment System

Measure Developer: Yale/CORE; Measure Steward: CMS

We thank Panel members for the time they took to provide their thoughtful comments and feedback. We have addressed your questions in the responses below.

Reliability

- **Issue 1:** Measure specifications, attribution: One reviewer was concerned about the complexity of the attribution process and noted that the developers did not support the use of visit acuity vs. visit frequency. The reviewer also inquired about the possibility of multiple attribution (specialists and PCP), and within the same group. The reviewer also asked about the stability of attribution to an individual provider.
    - **Developer Response 1:** We thank the reviewer for their comments. The attribution algorithm for the Heart Failure measure was closely aligned with the attribution algorithm for the MIPS Multiple Chronic Conditions (MCC) measure which assigns patients to a single clinician (unique NPI/TIN) who is most responsible for patient's care

based on the number and pattern of E&M visits. The measure does not consider visit acuity since acute visits may be related to the quality of care delivered. We agree that there are instances where patients are seeing multiple providers who impact the risk of admission. An underlying premise of our approach to attribution, which is supported by our TEP and Clinician Committee, is that ideally there is an individual clinician who is taking responsibility for managing and coordinating the care of a HF patient. In most cases, this will be a PCP or a cardiologist. The attribution reflects where the patient received the majority of care for the year during which performance is assessed; this may be stable for some, but not others. Given the intent of the measure, patient was assigned based on care received during the measurement year.

- **Issue 2:** Measure specifications, definition of exclusions. One reviewer requested a definition of the exclusion for patients who were in hospice, and patients who had no E&M visits.
    - **Developer Response 2:** The measure excludes patients who were in hospice at any time during the year prior to the measurement year or at the start of the measurement year. Patients enrolled in Medicare's hospice benefit are identified through the Medicare Enrollment Database.
    - Evaluation & Management (E&M) visits were defined using the codes shown in table 6 of the data dictionary.
- **Issue 3:** Measure specifications, inclusion criteria. One reviewer noted that the inclusion criteria for admissions with a heart failure diagnosis as not clearly stated.
    - **Developer Response 3:** The measure includes admissions for patients with a principial discharge diagnosis on the index claim of heart failure, or a principal discharge diagnosis of cardiomyopathy if without any other diagnosis (principal or secondary) of heart failure. The codes used to define these categories are in table 1 of the data dictionary.
- **Issue 4:** Measure specifications, level of measurement: One reviewer noted that the measure was submitted for endorsement for clinicians and groups of clinicians, but it was unclear which, if any, tests were conducted at the individual clinician level.
    - **Developer Response 4:** Model testing, including model fit and calibration, were based on the patient-level risk adjustment model and thus agnostic of assignment to clinician or clinician group. Variation in measure scores and reliability were tested at the TIN level. Under MIPS, clinicians can decide annually whether to report as individuals (NPI/TIN), as part of a group (TIN), or as both. The TIN level includes both solo clinicians (clinicians opting not to report with other clinicians under MIPS) and groups of clinicians who have chosen to report their quality under a common TIN. Therefore, testing results include both individual clinicians and clinician groups, consistent with how the MIPS program evaluates quality. Among TINs with at least 21 heart failure patients (minimum reliability of 0.4), 31.8% represent solo providers.
- **Issue 5:** Reliability, calculation and interpretation. One reviewer noted that they did not recognize the formula used to calculate the signal-to-noise reliability and asked for clarification on the interpretation of the result.
    - **Developer Response 5:** ICC in this context quantifies the proportion of variance explained by a grouping (random) factor in multilevel/hierarchical data. Yes, the

reliability estimate can still be characterized as the squared correlation between a measurement and the true value. To estimate the overall signal and noise, we first calculated the ICC for the provider entity (TIN) j using the estimates of between-entity variance σ^2, dispersion parameter ω, and mean of outcome λ, from a hierarchical generalized linear model (HGLM). The formula appropriate for the NB-1 model is $ICC_j = \sigma^2/(\sigma^2 + \ln(1+ \omega/\lambda))$. We then used the equation: $R_j = n_j ICC_j/(1+(n_j-1)ICC_j)$ where n_j is the number of observations for each entity, to calculate the reliability of each entity measurement. R_j can range from 0 (less than chance agreement) to 1.0 (perfect agreement).

- **Issue 6:** Reliability, clarification of unit of analysis: One reviewer asked for a clarification regarding the unit of analysis for the reliability distribution shown on page 8 and in section 2a.2.4.
    - **Developer Response 6:** The table on page 8 shows the distribution of reliability results for three levels of analysis in three separate columns shown in the table headings, from left to right: all TINs, TINs with a volume of >=21 patients, and TINs with a volume of >=32 patients. The results in section 2a.24 reflect results for TINs with at least 21 patients, as stated in that section: "Calculated using one year of data, the median measure score reliability was 0.600 for TINs with at least 21 patients, which is considered adequate [1, 2]."
- **Issue 7:** Reliability, description of providers: One panel member noted that the distribution of the sample appears to favor group practices, but it is not clear whether these are small/large group practices or what specialties are represented in these practices. The reviewer noted that reliability coefficients even at the higher level reported would still be of concern for between group comparisons.
    - **Developer Response 7:** We are not entirely sure what the reviewer refers to when they say that the distribution appears to favor group practices. On average, group practices have more patients and since patient volume is the driver of reliability, they tend to have higher reliability. About 31.8% of the TINs represent solo providers, 28.0% represent groups of 2-5 providers, 10.5% represent groups of 6-10 providers, and the remainder are practices of more than 10 providers. For determination of group size, the number of providers per group was based on all NPIs reporting under the same TIN, not just NPIs with attributed heart failure patients. Thus, all specialties were included in determining group size. A minimum case volume of 21 HF patients resulted in a median reliability of 0.6, IQR 0.48-0.78. Case volume was applied at the TIN level and did not take into account the number of patients per clinician when clinicians reported as a group under a single TIN.
- **Issue 8:** Reliability, minimum sample size: Several reviewers asked about the case-minimum requirements for the measure, and if the measure would be specified to be used with TINs with >21 patients, due to concerns about reliability.
    - **Developer Response 8:** CMS typically sets minimum volume of patients (and minimum number of providers, if applicable) during the process of rule-making. CMS intends to use the measure at the TIN level. In setting a minimum reliability threshold, CMS needs

to balance measure reliability with the statutory requirement to make performance measures applicable to the broadest number of providers. For these reasons, CMS typically sets a minimum reliability threshold of 0.4 across all providers, which will correspond to an average reliability of 0.636 in the MIPS program.

Validity

- **Issue 1:** One reviewer was concerned about the fact that the measure the only included 23.9% of provider groups and 69.8% of clinicians in the testing sample.
    - **Developer Response 1:** We note that the calculations of percent of TINs and percent of NPI/TINs were conducted on the sample of 45,093 TINs with at least 1 HF patient assigned. Since the measure is intended to distinguish the quality of care among MIPS clinician TINs that care for HF patients, it is not intended to apply to those TINs that care for just a few patients with HF (e.g., 1-2). When we raised the minimum patient volume to at least 21 patients, 10,760 or 23.9% of TINs were in this category. However, this represented 88.9% of patients with HF and 91.3% of the outcome; thus, the excluded TINs are TINs that participate in the care of very few eligible HF patients.
- **Issue 2:** Validity, face validity: One reviewer noted that for face validity, the panelists were asked about perceptions of the ability for the measure to discern quality at the group level but not at the individual clinician level. Another reviewer asked why only 12 of 17 TEP members voted.
    - **Developer Response 2:** We appreciate this comment. TEP members were asked whether the risk-standardized acute admission rates obtained from the HF measure as specified can be used to distinguish good from poor quality of care provided to HF patients by TINs reporting under MIPS. As noted above, TINs can be composed of clinician groups or individual clinician providers. Thus, validity was assessed as the measure is intended to be used (at the TIN level).

      Regarding the number of TEP members who responded, this was a multiyear TEP we convened for MIPS measure development, and some initial members were no longer participating by the end of the development.

- **Issue 3:** Validity, construct: One reviewer noted that heart failure patients are frequently hospitalized, and that the developer should have considered other important metrics, such as mortality.
    - **Developer Response 3:** We agree that mortality is an outcome of importance to patients with HF. In developing this measure, we focused on acute unplanned CV-related admissions because they represent an actionable subset of admissions that can be influenced by primary care providers (PCPs) and cardiologists. Acute CV-related admissions occur when outpatient management of HF fails, or when patients develop new or worsening symptoms or CV complications. The measure aims to incentivize effective and coordinated care for patients with HF to reduce the rates of these admissions.
    - As the reviewer notes, patients with HF are at high risk of both hospital admissions and mortality, and we have examined this during measure development. Patients who die in the measurement year tend to be admitted more often in that year. In our cohort, 5.8%

died during the performance year. The mean crude CV-related admission rate was 211.6 per 100 person-years among patients who died and 21.0 per 100 person-years among those who remained alive. A better score on the measure is achieved by helping patients stay alive and contribute to the denominator while avoiding hospitalization.

- **Issue 4:** Validity, exclusions: One reviewer wanted to know why CKD-5 is excluded because of nephrology care, but CKD-4 is not? CKD-4 is, by guideline, under nephrology care, not cardiology care.
    - **Developer Response 4:** We agree that patients with CKD-4 should receive care from a nephrologist. However, unlike patients with end stage renal disease who exclusively rely on nephrology care for fluid/volume management, this group represents patients who may still be actively managed by cardiologists or PCPs for their cardiovascular needs.
- **Issue 5:** Validity, risk model: One reviewer noted that the risk model lacks indicators of heart failure and was concerned about the ability of the measure to account for differences in case mix between primary care physicians and specialty cardiologists.
    - **Developer Response 5**: We appreciate this concern. The MIPS HF measure accounts for patients with more complicated or severe heart failure in several ways:
        A. Exclusion of patients at advanced stages of heart failure, such as those with implanted left ventricular assist device (LVAD), those who receive home inotropic therapy, or those with prior heart transplant or with end stage renal disease
        B. Risk adjustment for AICDs (defibrillators)
        C. Risk adjustment for systolic heart failure (which portends a poor prognosis)
        D. Risk adjustment for comorbidities including chronic kidney disease, and for frailty/disability.
    - Multiple panel members noted that calibration for the model is “good” or “adequate” suggesting that the model performance well across strata of patient risk.
    - Across the 45,093 TINs with at least one HF patient, RSCAR measure scores ranged from 9.6 to 62.4 per 100 person-years, with a median of 24.8 and an IQR of 24.0 to 25.9. Similar distribution in measure scores was found across cardiology TINs: median of 25.2 and an IQR of 23.4 to 25.2 with range of 12.0 to 50.7 per 100 person-years.
    - We are conducting additional analyses to understand adequacy of risk adjustment.
- **Issue 6:** Validity, risk model: One reviewer was concerned that the model does not account for the difference between a single patient with multiple admissions vs. multiple patients with a single admission.
    - **Developer Response 6:** We model the outcome at the unique patient level, with each patient having 0, 1, 2, etc. unplanned cardiovascular-related admissions during the measurement year. The number of admissions for each patient is associated with the patient’s comorbidities (which are accounted for in the model), and with provider’s quality of care. Since the model is at the patient level, we are treating the two scenarios mentioned differently and accordingly. Specifically, for a provider to have a better score (lower RSCAR), the provider needs to aim to reduce admissions among all of their patients, including patients at high risk for multiple events.
- **Issue 7:** Validity, risk model: One reviewer noted that response to medication can differ by race and wanted to know why the model was not adjusted for race.

- **Developer Response 7:** We tested and included age but not race, consistent with the rationale and approach taken for other CMS outcome measures. Studies suggest that race-based differences in outcomes are generally driven by age and comorbidities (which we include in the risk adjustment), as well as disparities in care delivery (for example, women with HF tend to receive less evidence-based treatment), and not by biological differences.

- **Issue 8:** Validity, c-statistic: One reviewer noted that the c-statistic of the model was not provided.
  - **Developer Response 8:** Due to the type of model that is used for this measure, we provide the deviance R^2 rather than a c-statistic. The deviance R-squared evaluates how successful the fit is in explaining the variation of the data and can take on any value between zero and one, with a value closer to one indicating that a greater proportion of deviance is accounted for by the model. For example, a deviance R-squared value of 0.12 means that the fit explains 12% of the total deviance.
- **Issue 9:** Meaningful differences: One reviewer requested that we provide the standard error of measurement. Another reviewer noted that submitting performance in percentiles was not responsive to the question, which was to identify "meaningful differences."
  - **Developer Response 9:** Standard error of the measure score can be calculated using bootstrapping. However, we did not perform bootstrapping since the MIPS program reports scores in deciles. We provided deciles of RSCARs to show variation across providers and calculated the median incidence rate ratio to describe meaningful differences.

Measure Number: 3622

Measure Title: National Core Indicators for Intellectual and Developmental Disabilities (ID/DD) Home- and Community-Based Services (HCBS) Measures

Measure Developer/Steward: Human Services Research Institute, The National Association of State Directors of Developmental Disabilities Services (NASDDDS)

Validity

- **Issue 1:** Confirmatory factor analysis results were not reported for multi-item scales.
  - Developer Response 1: We thank the panel members for noting the omission. We conducted confirmatory factor analysis to test the factor structure of the five multi-item measures:
    - Community Inclusion Scale (CI-4)
    - Satisfaction with Community Inclusion Scale (PCP-5)
    - Transportation Availability Scale (CI-3)
    - Life Decisions Scale (CC-4)
    - Respect for Personal Space Scale (HLR-1)
    - Results indicate that the estimated model fits the data reasonably well:
      - Tucker Lewis index (TLI) = 0.924 (≥ 0.90 is acceptable)
      - Comparative Fit Index (CFI) = 0.942 (≥ 0.90 is acceptable)

        - Root Mean Square Error of Approximation (RMSEA) = 0.026 (< 0.05 indicates good fit)
    - We will include the results in our full submission
- **Issue 2:** Using correlations to support validity.
    - **Developer Response 2:** We thank the panel members for their comments related to correlations. We identified four related issues. To address these issues, we added information below to (1) provide theoretical/hypothetical context for the reported Pearson correlation coefficients, (2) correlate measures with external data, (3) report complete correlation results with proper corrections and (4) provide information about #PCP-1 (Community Job Goal), #PCP-3 (ADL Goal), #CI-1 (Social Connectedness), and #CI-3 (Transportation Availability Scale).

      We articulated directional hypotheses for expected associations among measures and only tested those hypotheses. State-level socioeconomic measures derived from 2018 data provided by the Census Bureau were included in the hypothesis testing. All 14 measures were supported in at least one hypothesis.

      The table below lists our hypotheses and their test results. All hypotheses were directional; one-tailed tests were conducted.

Measure(s)	Relational Hypothesis	Test Results
Community Job Goal (PCP-1)	Urban settings provide a broader range of employment opportunities and hence, a larger choice of the types of jobs that are available for people with IDD. Urbanized states would be expected to find it easier to meet individuals' need for employment by including their wish for employment in their service plans.	Percentage of a state's population living in urban areas is positively and significantly correlated with PCP-1. r = 0.395 p = 0.011
Community Job Goal (PCP-1)	State Employment Leadership Network (SELN) was established in 2006 to support state public managers to offer expanded community-based employment options for people with IDD. We would expect SELN member states to have greater capacity and incentives to support the employment needs of their service participants and hence, to score higher on this measure.	Mean PCP-1 score for SELN member states is 0.4146 compared to a mean score of 0.3347 for non-member states. A one-tailed t-test of the difference between means yields a p-value of 0.055. Keeping in mind the low sample size (37 states), this result provides some support for the hypothesis.
Person-Centered Goals (PCP-2) and Chose Staff (CC-1)	State IDD service systems where it is common practice to develop a service plan based on the individual's preferences is likely to also provide staffing options based on their preferences.	These two measures are positively and significantly correlated. r = 0.344 p = 0.023

Measure(s)	Relational Hypothesis	Test Results
ADL Goal (PCP-3) and Community Job Goal (PCP-1)	In the support delivery system, Activities of Daily Living is a foundational element of assessment of functional support need, which is used to establish eligibility for service. Deficits in the ability to perform ADLs are therefore considered "low hanging fruit" in terms of developing a support plan. In many cases, ADL Goals may be carried over year over year for an adult receiving services, and it is unclear if they have chosen that goal. We would expect that service systems that use a deficit-based assessment and service planning may be more likely to have ADL goals in service plans. Person-centered-plans that are more progressive seek to support adults with IDD in their community employment and community participation goals, regardless of ADL deficits. We would expect to see community employment goals associated with a person's desire for community employment in progressive state service systems. We hypothesize a negative correlation between PCP-3 (ADL goal) and PCP-1 (Community job goal)	These two measures are negatively and significantly correlated. r = -0.342 p = 0.024
Lifelong Learning (PCP-4) and Has Friends (CI-2)	People with wider social circles are more likely to get exposed to new ideas and concepts. Therefore, states where people with IDD are more likely to report having friends (outside family and staff) would be expected to also have a high proportion of people with IDD reporting opportunities for lifelong learning.	These two measures are positively and significantly correlated. r = 0.764 p < 0.001
Satisfaction with Community Inclusion Scale (PCP-5) and Lifelong Learning (PCP-4)	Exposure to new ideas and concepts would be expected to increase one's expectations of inclusion in a broader range of community activities, thus increasing the sense of "relative deprivation" and increasing dissatisfaction with one's current level of community inclusion. We would therefore expect a negative association between PCP-5 and PCP-4	These two measures are negatively and significantly correlated. r = -0.498 p = 0.002

Measure(s)	Relational Hypothesis	Test Results
Social Connectedness (CI-1) and Respect for Personal Space (HLR-1)	People whose personal space is respected by the people with whom they interact are more likely to feel socially connected to them. We would therefore expect a positive association between social connectedness and respect for personal space	These two measures are positively and significantly correlated. r = 0.387 p = 0.012
Transportation Availability Scale (CI-3) and Satisfaction with Community Inclusion Scale (PCP-5)	People who have readily available means of transportation are more likely to be satisfied with their level of engagement in activities outside the home. CI-3 and PCP-5 would therefore be expected to be positively associated.	These two measures are positively and significantly correlated. r = 0.404 p = 0.009
Community Inclusion Scale (CI-4)	Resource-rich states would be expected to have greater ability to support people with IDD to engage in activities outside their home. Two measures of state-level resource availability were used to test this hypothesis: per-capita income and per-capita number of jobs.	Both measures of state-level resources are positively and significantly correlated with CI-4. Per-capita income: r = 0.345 p = 0.023 Per-capita number of jobs: r = 0.471 p = 0.003
Can Change Case Manager (CC-2) and Life Decisions Scale (CC-4)	A state system that allows its clients a high degree of choice in life decisions is expected to also allow them to choose or to change their case managers, given that these two areas of choice reflect a common service philosophy.	These two measures are positively and significantly correlated. r = 0.349 p = 0.022
Can Stay Home When Others Leave (CC-3) and Life Decisions Scale (CC-4)	A state system that allows its clients a high degree of choice in life decisions is expected to also provide them with the option of staying home alone when others leave, given that these two areas of choice reflect a common service philosophy.	These two measures are positively and significantly correlated. r = 0.552 p = 0.001

In conclusion, all 14 measures have an association with at least one other measure in line with theoretical expectations. In addition to expected associations with each other, hypothesized associations with measures based on external data are also supported. These findings provide evidence of validity at the measured entity level.

- **Issue 3:** There were concerns regarding whether or not the submitted measures should be considered PRO-PMs.

- **Developer Response 3:** We thank the panel members for the opportunity to discuss this important point. One panel member noted "… these are not outcome measures… these are things that are done or not done – processes of care", "Responses to some of the measures can be easily influenced by environmental factors … that are external to the support programs being measured". To address these concerns, we provide this further information for the panel members' re-consideration of the proposed measures as Person-Reported Outcome Performance Measures (PRO-PMs) in the context of Home and Community-Based Services (HCBS).
  In considering whether our proposed measures are PRO-PMs, we consulted the following definitions, featured in NQF publication *Patient Reported Outcomes (PROs) in Performance Measurement* **(**NQF, 2013):
  "Patient-reported outcomes (PROs) are defined as 'any report of the status of a patient's (or person's) health condition, health behavior, or experience with healthcare that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else.'" (p.5)
  "A PRO-based performance measure (PRO-PM) is based on PRO data aggregated for an entity deemed as accountable for the quality of care or services delivered. Such entities can include (but would not be limited to) long-term support services providers, hospitals, physician practices, or accountable care organizations (ACOs). NQF endorses PRO-PMs for purposes of performance improvement and accountability…" (p.5)
  In fact, in Table 1 (p.5), our project (National Core Indicators) was used as an example for PRO-PM, which affirms NQF's view on HCBS outcomes as PRO-PMs.
  The National Quality Forum (NQF) defines HCBS as "an array of services and supports delivered in the home or other integrated community setting that promote the independence, health and well-being, self-determination, and community inclusion of a person of any age who has significant, long-term physical, cognitive, sensory, and/or behavioral health needs" (NQF 2016). Measures of the quality of these services must therefore include outcomes within a broad range of life domains.
  In the context of quality monitoring, the proposed measures are using person-reported data to assess the extent to which people who are in receipt of funded services are experiencing quality life outcomes. In a discussion of person-reported outcomes in HCBS, Lipson (2019) points out that advances in the field of disability have broadened the understanding of quality of life and how it is measured from an individual's perspective. Quality domains central to quality of life include choice and satisfaction with residential settings, as well as addressing barriers to community participation such as limitations in transportation.
- To conclude, we contend that the submitted measures qualify as PRO-PMs in the context of HCBS. We acknowledge the panel member's comment that "responses to some of the measures can be easily influenced by environmental factors … that are external to the support programs being measured". For the measures we have put forward, however, we suggest that effective and flexible home and community-based supports can be developed to address environmental factors that may be serving as barriers. Measures that address individual choice and key life outcomes such as

employment and community access reflect on quality of HCBS and are reported at the person-level as personal outcomes. Performance measures that give credit to HCBS providers for successfully overcoming environmental barriers to independent living and community integration should be given consideration for inclusion in measurement systems.

- **Issue 4:** Panel members sought additional evidence of validity through connection to other measures of quality.
    - **Developer Response 4:** We thank the panel members for their comments and acknowledge that there is a lack of presentation of external evidence in our original MIF and testing form to support the validity of the submitted measures and instruments. Here we include additional information for consideration.
    - Measures from National Core Indicators have been cross-walked and tested for their applicability for benchmarking, assessing, quality monitoring and comparing progress at various levels and contexts. Below list some of the external evidence by state, national and international level.
    - **At the state level**

State	Evidence
Arizona	Increased provider rates to incentivize Community and Supported Employment initiatives; Created District Employment Specialist positions, showing that components of #PCP-1 is relevant in the state's policymaking. (Bradley, Hiersteiner, &Bonardi 2016)
Massachusetts	The state department of developmental services Licensure and Certification data, an external state-level data source, clearly corroborates with NCI measure #CI-1 Social Connectedness. This was referenced by the state in a brief analysis report: Quality is No Accident. https://shriver.umassmed.edu/wp-content/uploads/2020/07/QINA-Friendship_final_web2.pdf
Kentucky	In 2010, the Kentucky Division of Developmental and Intellectual Disabilities implemented changes related to NCI measure #PCP-1 Community Job Goal and #CI-1 Social Connectedness, showing the relevance of those measures.
Many states, such as Tennessee and Michigan	Convene committees and quality improvement councils to review NCI data, which includes the submitted NCI measures.

    - **At the National Level:** The National Core Indicators Measures selected for submission have demonstrated validity through alignment with multiple quality monitoring frameworks and tools, detailed below:
        - Medicaid Adult Core Health Care Quality Measure Set: In 2019, the Center for Medicaid and CHIP Services (CMCS) announced updates to the Medicaid Adult Core Health Care Quality Measure Set to include use of the National Core Indicators® (NCI®) to measure the quality of healthcare provided to adult

Medicaid recipients on three measures [Life decisions scale (#CC-4); Transportation measure (component of #CI-3); everyday choices scale (includes #CC-1 and #CC-2)] that were selected to be reported to CMS this year. The Adult Core Set Measures are available at https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-and-child-health-care-quality-measures/adult-health-care-quality-measures/index.html.

- Recommended Measure set for Medicaid-Funded Home and Community Based Services: In 2020, CMS proposed a measure set for quality monitoring. Each of the measures in this submission were included as part of the CMS-proposed measure set. Detailed information is available at the following link. HCBS Recommended Measure Set RFI (medicaid.gov)
- Medicaid Scorecard: NCI was one of the three experience-of-care surveys included in the Medicaid Scorecard for LTSS, which is used by CMS to increase public transparency and accountability about the state Medicaid programs' administration and outcomes. Details here: https://www.medicaid.gov/state-overviews/scorecard/state-use-patient-surveys-ltss-beneficiaries/index.html
- HCBS Advocacy Coalition's inclusion of NCI in Settings Rule Monitoring: The Medicaid Home and Community Based Services (HCBS) Settings Rule, issued by CMS in 2014, requires states to engage in ongoing monitoring throughout implementation. The #CI-4 Community Inclusion Scale was recommended in a white paper as a monitoring tool for this purpose. Further details are available at https://hcbsadvocacy.org/2020-outcomes-paper/.

- **At the International Level:** Measures from National Core Indicators have been cross-walked and tested for their applicability for benchmarking, assessing, and comparing progress towards a more inclusive society as described in the United Nations Convention on the Rights of People with Disabilities (UNCRPD), which includes a provision for the monitoring of outcomes of people with disability (Articles 31 and 33). To that end, National Core Indicators has been identified as providing a potential pathway to measurement of outcomes for people with IDD along key domains. Ticha et al (2018) laid out a conceptual framework for alignment of the UNCRPD with National Core Indicators. As a follow up study, Houseworth's analysis (Houseworth et al, 2019) was expanded to empirically test the framework and groupings of National Core Indicators by Articles of the UNCRPD. The results of our factor analysis largely aligned with Tichá et al.'s (2018) grouping of NCI items by UNCRPD article.
- In sum, all these work establish that the submitted measures have strong face validity, are widely recognized for relevance, and corroborated with externally-sourced data.

- **Issue 5:** Exclusion criteria were unclear and inconsistent on MIF and testing forms. One panel member noted that "…the MIF notes a number of exclusions. However, the testing form checks the box for 'no exclusions'…"
  - **Developer Response 5:** We acknowledge the misalignment between testing form section 2b2.(exclusion analysis) and MIF sections about exclusions(s.8, s.9, s16), and would like to clarify and provide additional information.

- Before we clarify about the exclusion criteria, some important context: To facilitate and accommodate person-centered reporting, the data collection instrument is divided into two sections, denoted by Roman numerals I and II. Section I of the survey contains questions about personal experiences and therefore may only be answered by the individual receiving developmental disabilities services. Section II of the survey---featuring questions about topics such as community involvement, choices, rights, and access to services—allows for responses from a "proxy," defined as a person who knows the individual well (such as a family member or friend).
- At the end of Section I, the surveyor assesses whether the respondent appears to understand at least one question and answers in a cohesive manner. This assessment is the only subjective process in the exclusion determination process, but it is not done on an arbitrary or state-by-state basis. Rather, it is based on a protocol, included in the manual and reviewed during surveyor trainings, that apply uniformly to all surveyors across different participating states. The protocol is straightforward—the section must be marked "valid" if at least one question in the section was answered in a manner that the basic level of comprehension was shown, and a clear response given either verbally (e.g. yes/no) or non-verbally (nodding/shaking head).
- NCI and participating states routinely conduct surveyor training and surveyor shadowing and reviewing processes that ensure, among other things, that surveyors are applying this assessment (whether or not Section I was valid) strictly based on the protocol.
- A proxy is not required, and sometimes no proxy is available, so the person with disabilities may answer both Section I and Section II (which is important for criteria c below).
- There are 4 section-based exclusion criteria:
  For Section I items:
  (a) Based on survey protocol, the surveyor found that the respondent did not give any valid responses to any Section I questions, or
  (b) All questions in Section I were left blank, or marked "not applicable" or "don't know".
  For Section II items:
  (c) Section II was completed without using a proxy, while Section I was deemed invalid (see criteria a above), or
  (d) All questions in Section II were left blank, or marked "not applicable" or "don't know".
  Here is the distribution of exclusions among states:

Exclusion criterion	N excluded	% excluded	Distribution across states N=22,009 (Min, 25th, 50th, 75th percentile, and max)
For Section I items: (a) Based on survey protocol, the surveyor found that the respondent did not give any valid responses to any Section I questions	5,053	22.9%	(3%, 13%, 23%, 26%, 62%)
(b) All questions in Section I were left blank, or marked "not applicable" or "don't know".	1,882	8.6%	(0%, 0%, 5%, 14%, 42%)
For Section II items: (c) Section II was completed without using proxy, while Section I was deemed invalid (see criteria a above), or	59	0.3%	(0%, 0%, 0%, 0%, 5%)
(d) All questions in Section II were left blank, or marked "not applicable" or "don't know". Here is the distribution of exclusions among states	311	1.4%	(0%, 0%, 1%, 1%, 7%)

Interpretations:

- Exclusion (a), (the surveyor found that the respondent did not give any valid responses to any Section I questions), accounting for 22.9% of all surveys, represents the majority (69%) of all exclusions and is meant to safeguard the validity of measures that utilize Section I items. This exclusion, with its conservative approach, prevents the inclusion of responses with sub-standard reliability in measure calculations.
- Exclusion (b) (All questions in Section I were left blank or marked "not applicable" or "don't know") accounts for about 8.6% of all surveys and represents about a quarter (26%) of all exclusions. It is purely objective and is needed to prevent the inclusion of responses that do not contribute meaningful data for Section I items.
- Exclusion (c) (Section II was completed without using proxy, while Section I was deemed invalid) only accounts for 0.3% of all surveys and represents less than 1% of all exclusions. However, it is in place to safeguard the validity and reliability of measures that utilize Section II items by excluding responses provided by an individual whose responses to Section I were assessed as unreliable. Given that a very small percentage of individuals are being excluded, it is unlikely this exclusion unduly affects the measure score.
- Exclusion (d) (All questions in Section II were left blank or marked "not applicable" or "don't know") accounts for 1.4% of all surveys, 4% of all exclusions. Its determination is purely objective and is needed to prevent the inclusion of surveys that do not contribute

meaningful data for Section I items.

- One panel member noted that answers of “unknown” or “not applicable” are dropped from the denominator. We thank the panel member for noting the lack of clarity on this and will amend our numerator and denominator statements in the full submission. Answers of “unknown” or “not applicable” do not get included because including such answers in the calculations would cause underestimation—for example, those who already have a job would not have a job-related goal in their HCBS service plan, and would answer “not applicable” to “would you like to have a job in the community”. By including such answers, the #PCP-1 Community Job Goal calculations would have lower rates than otherwise, thereby masking the true gap (those who want a job but do not have a job goal) in quality monitoring.
- In conclusion, exclusions are based on the uniformly applied criteria, most of which are objective and all of which are standardized. The exclusions were put in place to ensure accurate calculation of the measures and to safeguard validity and reliability. It is important to note that to the extent possible, exclusions eliminate unreliable responses, not the entire survey. For example, a survey where Section I responses are excluded from measure calculations may still be included in measures based on Section II items if Section II responses were provided by a proxy. We intend to amend the testing attachment section 2b2 to align with MIF s.8, s.9 and s.16.

- **Issue 6:** States cherry-pick favorable surveys or survey sites.
  - **Developer Response 6:** States’ survey strategies are determined by workplans, which are third-party designed and reviewed. Many states contract with surveying agencies to conduct surveys. States do not get to pick “successful” sites or programs for interviewing. The National Association of State Directors of Developmental Disabilities Services (NASDDDS) provides general oversight and guidance in all states NCI activities. All state HCBS eligible populations are generally included in the survey frame, unless reasonable justifications can be made.

References

1. Bradley, V.J., Hiersteiner, D., Bonardi, A. (2016). A focus on System-Level Outcome Indicators. In Cross Cultural Quality of Life: Enhancing the Lives of People with Disability. Schalock and Keith Eds. American Association on Intellectual and Developmental Disability.
2. Houseworth, J., Stancliffe, R., & Tichá, R. (2019). Examining the National Core Indicators' Potential to Monitor Rights of People with Intellectual and Developmental Disabilities According to the CRPD. Journal of Policy and Practice in Intellectual Disabilities, 16(4), 342-351
3. Lipson D.J. (2019), Person-Reported Outcome Measures for Home and Community-Based Services. HCBS Quality Measures Issue Brief. Mathematica https://www.medicaid.gov/medicaid/quality-of-care/downloads/hcbs-quality-measures-brief-2-person-reported-outcome.pdf
4. National Quality Forum (2013). Patient Reported Outcomes (PROs) in Performance Measurement. Final Report. Washington, DC: NQF, January 2013. Available at https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=72537
5. National Quality Forum (2016). Quality in Home- and Community-Based Services to Support Community Living: Addressing Gaps in Performance Measurement. Final Report. Washington, DC:

NQF, September 2016. Available at https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=83433
6. Tichá, R., Qian, X., Stancliffe, R.J., Larson, S., & Bonardi, A. (2018). Alignment between the Convention on the Rights of Persons with Disabilities and the National Core Indicators. Journal of Policy and Practice in Intellectual Disabilities, 15(3), 247-255.

Subgroup 2

Measure Number: 2431

Measure Title: Hospital‑level, risk‑standardized payment associated with a 30‑day episode-of-care for Acute Myocardial Infarction (AMI)

Measure Developer/Steward: Developer: CORE/Steward: CMS

We thank the Scientific Methods Panel members for their thoughtful comments and for the time they took to review the measure specifications and testing results. Below we provide responses to questions posed by the Panel members within their Preliminary Assessment.

Reliability

- **Issue 1:** A Panel member asked if, in the case where a patient is transferred from an emergency department to another hospital and then subsequently admitted, if CORE had examined if the emergency department is affiliated with a hospital without capability to treat the AMI patient or whether the hospital simply does not want to take the risk of admitting the patient?
    - **Developer Response 1:** We have not looked at this particular scenario, but the payment measures are aligned with the methodology of the 30 day-mortality wherein the outcome is attributed to the first hospital that admits the patient.

Validity

- **Issue 1:** One Panel member expressed concern over not using social risk factors such as dual eligibility status in the model.
    - **Developer Response 1:** The payment measures are meant to be reported along with readmission and mortality measures for the same conditions, and those measures, which were recently recommended for re-endorsement, do not include adjustment for social risk factors.
    - We do not dispute that there are differences in unadjusted, observed outcomes based on social risk - our own results presented in the testing attachment show for example, that for the dual eligibility variable, mean observed payments for AMI are higher for patients with the social risk factor compared with patients without the social risk factor. (Note that due to feedback from NQF, we did not test any race-related variables.)
    - The question we are trying to address with our analyses is the impact of adjusting for social risk factor on this particular measure score (risk-standardized payment). Our results show that differences in mean payments are very small, and the correlations between adjusted and unadjusted RSPs are near 1. In addition, our approach is consistent with recommendations from the Office of the Assistant Secretary of Planning and Evaluation (ASPE) that resource measures used in public reporting should not be adjusted for social risk.[1]

- **Issue 2:** One Panel member expressed concern that Medicare Advantage data was not included in the measure.
    - **Developer Response 2:** CMS is actively investigating the reliability and validity of Medicare Advantage data for inclusion in measures.
- **Issue 3:** One Panel member expressed concern about the exclusion "Discharged alive on the day of admission or the following day who were not transferred to another facility" because the underlying assumption is that the patient discharged within one day of hospitalization may not have clinically significant AMI. The Panel member also asked for the typical or average length of hospitalization for AMI?
    - **Developer Response 3:** A patient with a clinically significant AMI would likely have had a subsequent inpatient stay and not have been discharged on the same day or the following day. The patient may have been held for observation, and then released. This issue was brought up during the NQF re-endorsement evaluation of the AMI mortality measure, and we will be examining this in more detail in an upcoming annual measure review cycle. According to a 2020 study, mean length of stay was 5 days.[2]
- **Issue 4:** One Panel member asked about the use of ICD-9 codes in the submission and wondered why they would be used since the transition to ICD-10 has already occurred.
    - **Developer Response 4:** ICD-9 codes were mentioned (in purple text to indicate information from the prior submission) as they were used during measure development, when ICD-9 codes were still in use. The measure currently uses ICD-10 codes as described in the data dictionary. To convert measures from ICD-9 to ICD-10, CORE implements an extensive code mapping, review, and measure validation process to ensure the validity of the updated measure.
- **Issue 5:** One Panel member noted that standardized pricing models can mask real differences in resources available across hospitals due to payer mix and asked if the measure used standardized pricing**.**
    - **Developer Response 5:** The goal of the measure is to compare resource use for this specific condition among Medicare Fee-for-Service (FFS) beneficiaries, therefore payer mix should not mask differences in resources used to treat FFS beneficiaries. The measure aims to identify resources used, not resources that are available. The measure does use price standardized reimbursements.

References

1. 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/pdf-report/second-impact-report-to-congress. Accessed January 4, 2021.
2. Wang, Y., Leifheit, E., Normand, S. T., & Krumholz, H. M. (2020). Association Between Subsequent Hospitalizations and Recurrent Acute Myocardial Infarction Within 1 Year After Acute Myocardial Infarction. Journal of the American Heart Association, 9(6), e014907. https://doi.org/10.1161/JAHA.119.014907

Measure Number: 2436

Measure Title: Hospital-level, risk-standardized payment associated with a 30-day episode-of-care for heart failure

Measure Developer: Yale CORE; Steward: CMS

We thank the Scientific Methods Panel members for their thoughtful comments and for the time they took to review the measure specifications and testing results. Below we provide responses to questions posed by the Panel members within their Preliminary Assessment.

Validity

- **Issue 1:** One Panel member expressed concern over not using social risk factors such as dual eligibility status in the model.
    - **Developer Response 1:** The payment measures are meant to be reported along with readmission and mortality measures for the same conditions, and those measures, which were recently recommended for re-endorsement, do not include adjustment for social risk factors.
    - We do not dispute that there are differences in unadjusted, observed outcomes based on social risk - our own results presented in the testing attachment show for example, that for the dual eligibility variable, mean observed payments for heart failure are somewhat higher for patients with the social risk factor compared with patients without the social risk factor, but for dual eligibility, payments are somewhat lower for patients with the social risk factor. (Note that due to feedback from NQF, we did not test any race-related variables.)
    - The question we are trying to address with our analyses is the impact of adjusting for social risk factor on this particular measure score (risk-standardized payment). Our results show that differences in mean payments are very small, and the correlations between adjusted and unadjusted RSPs are near 1. In addition, our approach is consistent with recommendations from the Office of the Assistant Secretary of Planning and Evaluation (ASPE) that resource measures used in public reporting should not be adjusted for social risk.[1]
    - In addition, adjusting for social risk factors would likely remove an important hospital level effect. A 2019 study, described in the testing attachment and authored by the developer, showed that differences in hospital-level payments for heart failure and pnumonia were associated with hospital characteristics independently from patient characteristics.[2] This study was able to hold constant the social determinants of health that were not expected to change between the two admissions; the study compared the same people at two different hospitals so that behaviors, social context, and demographic characteristics, including race/ethnicity, were the same. In this study, we compared payments for the same Medicare patient for two admissions for the same condition – one admission to a low-payment hospital and one admission to a high-payment hospital and found that patients who were admitted to hospitals with the highest payment profiles incurred higher costs than when they were admitted to hospitals with the lowest payment profiles. The findings suggest that variations in payments to hospitals are, at least in part, associated with the hospitals independently of non–time-varying patient characteristics.

- **Issue 2**: One Panel member expressed concern that Medicare Advantage data was not included in the measure.
    - **Developer Response 2**: CMS is actively investigating the reliability and validity of Medicare Advantage data for inclusion in measures.
- **Issue 3:** One Panel member asked about the use of ICD-9 codes in the submission and wondered why they would be used since the transition to ICD-10 has already occurred.
    - **Developer Response 3:** ICD-9 codes were mentioned (in purple text to indicate information from the prior submission) as they were used during measure development, when ICD-9 codes were still in use. The measure currently uses ICD-10 codes as described in the data dictionary. To convert measures from ICD-9 to ICD-10, CORE implements an extensive code mapping, review, and measure validation process to ensure the validity of the updated measure.
- **Issue 4:** One Panel member noted that standardized pricing models can mask real differences in resources available across hospitals due to payer mix and asked if the measure used standardized pricing**.**
    - **Developer Response 4:** The goal of the measure is to compare resource use for this specific condition among Medicare Fee-for-Service (FFS) beneficiaries, therefore payer mix should not mask differences in resources used to treat FFS beneficiaries. The measure aims to identify resources used, not resources that are available. The measure does use price standardized reimbursements.
- **Issue 5:** One Panel member expressed concern about the quasi-R-square value, noting that it was low. Another Panel member noted that the value is similar to other cost measures.
    - **Developer response 5:** That is, the R2 reflect the proportion of variance accounted for by patients' clinical risk factors. This suggests that payment is driven less by patient characteristics than other characteristics, such as attributes of the facility.

References

1. 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/pdf-report/second-impact-report-to-congress. Accessed January 4, 2021.
2. Krumholz, H. M., Wang, Y., Wang, K., Lin, Z., Bernheim, S. M., Xu, X., Desai, N. R., & Normand, S.T. 2019. Association of Hospital Payment Profiles With Variation in 30-Day Medicare Cost for Inpatients With Heart Failure or Pneumonia. JAMA network open, 2(11), e1915604.

Measure Number: 2579

Measure Title: Hospital‑level, risk‑standardized payment associated with a 30‑day episode-of-care for pneumonia

Measure Developer: Yale CORE; Steward: CMS

Reliability

- **Issue 1:** A Panel member asked for clarification of the specifications and if sepsis patients are included in the measure
    - **Developer Response 1:** Yes, sepsis patients are included in the measure. The inclusion

criteria are a principal diagnosis of:
    - Pneumonia; or,
    - Sepsis (not including severe sepsis) with a secondary diagnosis of pneumonia coded as present on admission (POA) and no secondary diagnosis of severe sepsis coded as POA.
- **Issue 2:** A Panel member asked for the ICD-10 codes for the inclusion criteria for pneumonia and sepsis, and a map to the DRGs included in the dataset.
    - **Developer Response 2**: The ICD-10 codes can be found in the data dictionary (Tab 1); There is no map for DRGs. It is proprietary software that is used by CMS for payment purposes. This measure uses the DRG that is assigned in the claim reconciliation process.

Validity

- **Issue 1:** One Panel member expressed concern over not using social risk factors such as dual eligibility status in the model.
    - **Developer Response 1:** The payment measures are meant to be reported along with readmission and mortality measures for the same conditions, and those measures, which were recently recommended for re-endorsement, do not include adjustment for social risk factors.
    - We do not dispute that there are differences in unadjusted, observed outcomes based on social risk - our own results presented in the testing attachment show for example, that for the dual eligibility variable, mean observed payments for pneumonia are higher for patients with the social risk factor compared with patients without the social risk factor. (Note that due to feedback from NQF, we did not test any race-related variables.)
    - The question we are trying to address with our analyses is the impact of adjusting for social risk factor on this particular measure score (risk-standardized payment). Our results show that differences in mean payments are very small, and the correlations between adjusted and unadjusted RSPs are near 1. In addition, our approach is consistent with recommendations from the Office of the Assistant Secretary of Planning and Evaluation (ASPE) that resource measures used in public reporting should not be adjusted for social risk.[1]
    - In addition, adjusting for social risk factors would likely remove an important hospital level effect. A 2019 study, described in the testing attachment and authored by the developer showed that differences in hospital-level payments for heart failure and pneumonia were associated with hospital characteristics independently from patient characteristics.[2] This study was able to hold constant the social determinants of health that were not expected to change between the two admissions; the study compared the same people at two different hospitals so that behaviors, social context, and demographic characteristics, including race/ethnicity, were the same. In this study, we compared payments for the same Medicare patient for two admissions for the same condition – one admission to a low-payment hospital and one admission to a high-payment hospital and found that patients who were admitted to hospitals with the highest payment profiles incurred higher costs than when they were admitted to hospitals with the lowest payment profiles. The findings suggest that that variations in payments to hospitals are, at least in part, associated with the hospitals independently of non–time-varying patient characteristics.

- **Issue 2:** One Panel member expressed concern that Medicare Advantage data was not included in the measure.
    - **Developer Response 2:** CMS is actively investigating the reliability and validity of Medicare Advantage data for inclusion in measures**.**
- **Issue 3:** One Panel member asked about the use of ICD-9 codes in the submission and wondered why they would be used since the transition to ICD-10 has already occurred.
    - **Developer Response 3:** ICD-9 codes were mentioned (in purple text to indicate information from the prior submission) as they were used during measure development, when ICD-9 codes were still in use. The measure currently uses ICD-10 codes as described in the data dictionary. To convert measures from ICD-9 to ICD-10, CORE implements an extensive code mapping, review, and measure validation process to ensure the validity of the updated measure.
- **Issue 4: One Panel member noted that** standardized pricing models can mask real differences in resources available across hospitals due to payer mix and asked if the measure used standardized pricing**.**
    - **Developer Response 4:** The goal of the measure is to compare resource use for this specific condition among Medicare Fee-for-Service (FFS) beneficiaries, therefore payer mix should not mask difference in resources use to treat FFS beneficiaries. The measure aims to identify resources used, not resources that are available. The measure does use price standardized reimbursements.

References

1. 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/pdf-report/second-impact-report-to-congress. Accessed January 4, 2021.
2. Krumholz, H. M., Wang, Y., Wang, K., Lin, Z., Bernheim, S. M., Xu, X., Desai, N. R., & Normand, S.T. 2019. Association of Hospital Payment Profiles With Variation in 30-Day Medicare Cost for Inpatients With Heart Failure or Pneumonia. JAMA network open, 2(11), e1915604.

Measure Number: 3614

Measure Title: Hospitalization After Release with Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke)

Measure Developer/Steward: Johns Hopkins Armstrong Institute for Patient Safety and Quality

Validity

DATA ELEMENT VALIDITY TESTING

- **Issue 1:** One reviewer noted the kappa statistic would be helpful to understand the chance corrected agreement rates between the data elements.
    - **Response:** The short answer is the Kappa statistics were >0.94. Additional detail on their calculation is provided below:
    - For the 'dizzy out' code validation (one year of data in the ICD-9 era, and one year of data in the ICD-10 era), there was an initial coding component performed by electronic

validation and a secondary coding component by human raters (for charts at higher risk of misclassification).

- ***For "positive" charts (i.e., discharged from the ED with a benign dizziness diagnosis ICD code):*** Results were first compared to the electronic health record chief complaint for that visit; if the chief complaint was "dizziness" (n=2,360) then the discharge ICD code was presumed valid. From the remaining "positive" charts (in which the dizziness discharge diagnosis did **not** match the chief complaint), we sampled 128 charts for manual review (drawing them at random, but deliberately oversampling half of these from the group of charts most likely to be erroneously labeled 'dizzy' – e.g., charts of those with hearing loss as a chief complaint, that might have been more likely to mention dizziness incidentally, thereby becoming mislabeled by chart coders). All the manually reviewed "positive" charts were deemed to be true positives (PPV 100%).
- ***For "negative" charts (i.e., NOT discharged from the ED with a benign dizziness diagnosis ICD code):*** We validated 160,966 charts as having something OTHER than dizziness as a chief complaint (i.e., non-dizzy ED discharge associated with non-dizzy ED chief complaint). From the remaining "negative" charts (in which the NOT dizziness discharge diagnosis yet **was** associated with a dizziness or other otologic chief complaint), we randomly sampled 134 charts for manual review. All but six of these charts were deemed to be true negatives (NPV 95.5% [human review of high-risk charts only]; NPV 99.997% [weighted average, all charts]).

  - Each of the 262 manually-reviewed charts were reviewed by two of four raters (#1-#4). The Cohen's Kappa statistic between Reviewer #1 and Reviewer #2 (n=129 charts) was 0.94 (95% CI is [0.89, 1]); the Kappa between Reviewer #3 and Reviewer #4 (n=133 charts) was 0.96 (95% CI is [0.91, 1]). Since these manual chart reviews were oversampled for "challenging" charts, the true Kappas for correct classification of the results for randomly selected charts is probably somewhere closer to 1.0.

- **Issue 2:** One reviewer noted concerns that the data element validity testing was conducted with only four hospitals from the same health system.
  - **Response:** The 'dizzy out' data validity testing was limited to one health system's EHR, but included a mix of both academic medical centers and community hospitals. The time periods used for testing reflected a time where each hospital in the health system used their own set of medical coders (they currently use a pooled set of coders, but did not during the time window of analysis presented in the application), so while all hospitals are part of the same health system, the coding was done independently across hospitals. In addition, given the strong positive feedback we received on our data element validity testing approach from the prior SMP review, we did not interpret that feedback to mean we needed to expand our data element validity testing to additional hospitals. But we are happy to explore opportunities for expansion in the future to ensure our findings are consistent across hospitals and health systems.
  - It is also important to remember that accuracy of coding an ED dizziness visit treat-and-release discharge is not likely to vary across hospitals or health systems. First, there are very few codes for dizziness and vertigo, compared to many other conditions, so there are few options to choose from (and they are all part of the same denominator set we

are using). Second, coding accuracy here is not a function of diagnostic accuracy (merely administrative billing accuracy for coding this as the intended ED visit diagnosis). Our data confirm the remarkable accuracy of such coding --- 100% with a lower 95% confidence bound of 99.9% on positive predictive value and 99.997% on negative predictive value with a lower 95% confidence bound of 99.99% on negative predictive value.

- **Issue 3: One reviewer noted that the** Tirchwell et al, and McCormick et al. studies on which we base many of our arguments with regard to data element validity only used ICD-9 codes and sought clarification on whether these findings automatically extended to ICD10 codes.
    - **Response:** Our submission included a reference to one study that compared the accuracy of ICD-9 and ICD-10 in capturing strokes (Kokotailo and Hill, 2005, *Stroke*). That study, conducted in three Canadian hospitals, found that stroke coding was equally good with ICD-9 (90% correct [95% CI 86-93]) as with ICD-10 (92% correct [95% CI 88-95]). Since our original submission, we have identified several additional studies that support ICD-10 as being a valid taxonomy for capturing strokes from administrative data:
        - (Chang et al.; 2019; *Stroke*): This study identified a small and transient decline in concordance between ICD-CM codes and stroke clinical diagnoses during the ICD-9/ICD-10 coding transition, indicating no substantial impact on the overall identification of stroke patients.
        - (Hsieh et al; 2021; *Clinical Epidemiology*): Using Taiwan's National Health Insurance claims database, the authors found among 983 hospitalizations, 860, 111, and 12 were determined to be true-positive, false-positive, and false-negative episodes of acute hemorrhagic stroke, respectively. The PPV and sensitivity of the ICD-10-CM codes of I60 or I61 for identifying acute hemorrhagic stroke were 88.6% and 98.6%, respectively.
        - (Hsieh et al; 2021; *Clinical Epidemiology*): Using Taiwan's National Health Insurance claims database, the authors found using ICD-10-CM code of I63 in any position of the discharge diagnoses to identify Acute Ischemic Stroke (AIS) yielded a PPV and sensitivity of 92.7% and 99.4%, respectively. The PPV increased to 99.8% with >12% decrease in the sensitivity when AIS was restricted to those with I63 as the primary diagnosis.

RISK ADJUSTMENT

- **Issue 4:** One reviewer noted that the same patient can be in the denominator multiple times, and his/her outcomes will be correlated. They raised the concern that the risk difference method used does not seem to have accounted for such correlations.
    - **Response:** We agree that these outcomes for a patient counted more than once may well be correlated, but we do not believe that such correlations should be accounted for through some form of statistical risk adjustment as part of the measure.
    - We allowed replacement up to three times for patients in a 3-year window. We did not allow replacement during the 360-day post-ED index visit period for calculating the expected (long-term baseline) rate, because this would have complicated whether a stroke event was counted as "observed" (short-term, acute) or "expected" (long-term, post-acute baseline). With the measure as constructed, ~95% of patients contribute only

one ED index visit in a 3-year period.
    - We made the decision to allow "replacement" of patients in the data set because we could not justify why someone misdiagnosed on multiple occasions should be considered "ineligible" to be misdiagnosed after the first time their stroke was missed. If anything, someone whose stroke was missed previously should be a red flag for a potential missed stroke on a subsequent ED visit --- accordingly, the subsequent miss would be even more clinically egregious than the first miss.
    - Failure to replace patients would also devalue (and make less measurable) the potential impact of patient-specific characteristics on misdiagnosis risk such as race, gender, or mental illness (all known risk factors for misdiagnosis). Imagine an African American female with dizziness caused by stroke is misdiagnosed three times more often than her white, male counterpart. If she is then misdiagnosed three times in a three year period (each time suffering a resulting stroke hospitalization and harm), while the white male is misdiagnosed once with the same outcome, why should each patient be considered to contribute only one misdiagnosis to the calculation? Our belief is "replacement" is important to capture this variation.
    - Finally, disallowing replacement would reduce the number of denominator and numerator events, reducing the measure's precision without a clear benefit.
- **Issue 5:** One reviewer noted that the measure states no risk adjustment, but is using an observed - expected rate to measure performance. In addition, there was lack of clarity on the definition of the expected rate.
    - **Response:** The expected rate is the stable, post-acute, long-term stroke rate in the same cohort (i.e., the exact same patients, not a matched cohort); this is a simple surrogate for the baseline stroke risk in that cohort, since it is well known that the untreated natural history of major stroke after minor stroke and TIA is that risk of major stroke, after an initial 'bump' returns to roughly the baseline rate (reviewed in Rothwell & Johnston, *Lancet Neurology*, 2006).
    - The guidance from NQF staff in prior discussions has been to *not* refer to this as a risk-adjustment model, as it does not include a set of "risk factors" that you would see in traditional risk-adjustment models. Instead, our approach to control for patient-level risk factors is to compare the short-term (acute, elevated risk) rate of a patient having a stroke after being discharged from the emergency department (within 30 days of discharge; the "observed" rate), as compared to the longer-term (post-acute, stable) baseline risk in that same patient population (within 90-360 days after discharge; the "expected" rate). In short, it is the same patient population being compared at two different biologically-relevant, disease natural history-defined differential time points within a 1-year time frame.
- **Issue 6:** One reviewer expressed concern that the "risk difference approach" may not sufficiently capture the "expected rate" of stroke, or that subtracting the calculated expected rate fully accounts for risk factors of patients. They also noted the lack of social risk adjustment was mentioned only in context of not "risk adjusting away" worse care for racial minorities, but there was no discussion of potential conceptual relationships (which may or may not exist).
    - **Response:** The rationale and details describing the 'expected' rate of stroke for the measured cohort are described in response to the prior reviewer's comment. The risk difference approach attempts to measure the observed (short-term, acute) greater than expected (baseline) risk of a stroke as a measure of a missed stroke diagnosis. As this

risk approach is comparing risk in the same patient population at two different time points, there is no adjustment for any patient-level risk factors, per se, including social risk factors. The assumption is any risk factors that are present in the "short-term/observed" (0-30 days post-ED discharge) for a patient are the same risk factors in place in the "long-term/expected" (90-360 days post-ED discharge) for the same patient. While it is certainly true that an individual's baseline stroke risk may vary over time (e.g., in response to disease progression or therapy --- or worsening or improvement of social circumstances), there is no reason to believe that, at a population level, this risk is anything other than stable over the subsequent 12 months. Multiple analyses from multiple countries using multiple different data sets have found the exact same return to a stable incidence rate (i.e., linear growth in cumulative incidence) for stroke after 90 days seen in these analyses (Kim et al, 2011, *Acad Emerg Med*; Atzema et al, 2015, *Ann Neurol*; Mane et al, 2018, *BMJ Qual Saf*; Chang et al, DEM Conference, 2017), mirroring exactly the known natural history of stroke risk over time following an acute minor stroke or TIA.

- As to the issue of conceptual relationships between risk of stroke and risk of misdiagnosis, it is absolute possible there are interactions. For example, we have previously shown using essentially the same method as submitted here that minority populations are at greater risk of being misdiagnosed (Mane et al, 2018, *BMJ Qual Saf*), and it is known that some minority populations are at higher risk of stroke (Morgenstern, 2004, *Am J Epi*). The risk difference approach accounts for the second, but not the first, allowing us to tease out the effect of race on misdiagnosis risk.

MEANINGFUL VARIATION IN PERFORMANCE

- **Issue 7:** One reviewer expressed concern that the 3-year performance period is quite long, and by the measure's construction may only apply to large hospitals with EDs discharging >40K ED patients yearly. They noted in an ideal world, the most potent way for validating this measure would have been that the measure developer to verify their estimated measure with the actual numbers from the facilities included in the final sample.
  - **Response:** The 3-year rolling window performance period is quite long, but it is very typical for low-frequency, important outcome measures (i.e., the CMS 30-day hospital mortality measures). The performance period could be shortened with access to more complete data sources. This is because the precision of the proposed measure is primarily a function of the number of outcome events (return strokes); this, in turn, is the product of the ED visit volumes and the event rate. The optimal data set for widespread, standardized measurement does not yet exist in the US. We chose Medicare FFS data because they are generally standardized in their coding across states and institutions, and they are highly valid in terms of tracking patients nationally (across hospitals, health systems, and states). However, Medicare data represent only a subset (~20%) of ED dizziness visits at most institutions. Thus, the number of outcome events is reduced by roughly 5-fold relative to the actual patient volumes encountered at each ED.
  - We have conducted some sensitivity analyses to understand whether hospitals could analyze their own hospital ED data with sufficient precision for a shorter (6-12 month) outcome measure. Although individual hospitals might not have access to hospital crossover data (i.e., discharged from Hospital A with dizziness and admitted to Hospital B with stroke), our analyses suggest that this may not be as much of a problem as one

might imagine. Hospital crossover rates are relatively high (e.g., 35-45%), so there is an important systematic bias towards under-counting of these events across all institutions. However, leaving out hospital crossover data from Medicare does not change the relative institutional performance ranking on the measure by more than +/- 1 decile for the majority of hospitals. Thus, hospitals could be benchmarked reasonably well against one another (particularly for identifying very low or very high outliers) just using self-reported data.

- For initial measure endorsement, NQF offers measure developers the opportunity to provide either data element validity testing or measure score validity testing. We choose to focus on data element validity testing for this submission. We agree that the two forms of testing do provide different insights on the measure's validity and we hope to be able to expand our testing in the future to assess the measure score's validity by comparing results to other measures of the same construct.

- **Issue 8:** A few reviewers expressed concerns with the skew in the final scores, with very few hospitals being labeled as a poor performer.
  - **Response:** When using just the Medicare FFS data (representing only ~20% of ED visits), just 8 of 927 facilities had a measure result in which the lower 95% confidence bound was greater than zero (i.e., "proving" that patients were being harmed by missed stroke events at rates above that expected by chance and baseline risk alone). However, there are many more hospitals than 8 with meaningfully worse performance that could be statistically demonstrated with a more complete reporting of ED visit data. As described above, the measure's precision could be readily enhanced using data from the individual hospitals (100% of ED visits, rather than ~20% contained in Medicare data). With greater precision comes greater resolving power for the measure, and greater ability to discriminate quality performance into groups or identify outliers with better or worse performance.
  - As shown in Figure 1 below, lower-volume hospitals are at higher risk of poor diagnostic performance than higher-volume hospitals (even when only analyzing the highest-volume hospitals, as was the case in our application). Precision at the low end of institutional visit volumes would be enhanced dramatically by including 100% of ED visits. This shift would allow far greater resolving power to identify performance gaps.
  - Thus, we do not believe that the issue here is one of measure validity or even precision of the measure itself, as articulated --- the real issue is the lack of an appropriate national data source to capture these events. This problem of incomplete data sets could easily be rectified if the measure were endorsed and required for public quality reporting purposes or benchmarking.

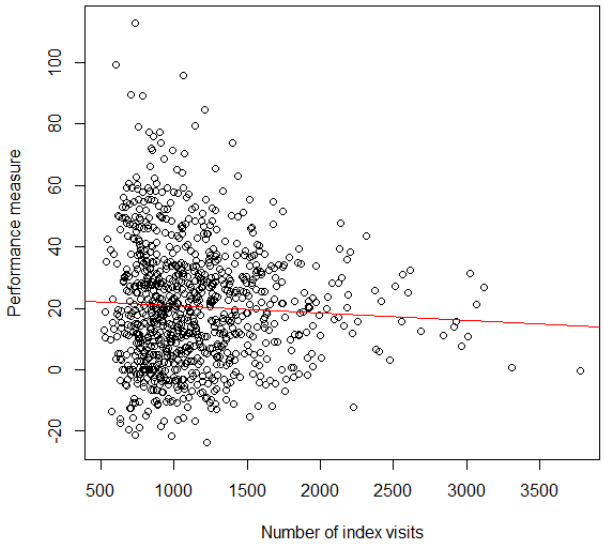


Figure 1. Relationship between ED index visit volume and performance measure (observed minus expected stroke hospitalizations after 'benign' dizziness discharge). ***Each circle represents a single institution in the analysis submitted to NQF.***

OTHER CONCERNS

- **Issue 9:** One reviewer expressed concern that the definition of the index ED visit excludes other ED visits within 360 days and that the impact of this exclusion was not tested.
    - **Response:** We are unclear to what exactly we would assess, but we are open to suggestions. If we had not excluded ED visits during the 360-day follow-up period (needed to calculate a stable long-term stroke 'baseline'), then a single stroke hospitalization could have counted as either 'observed' or 'expected' (or both).
    - In examining 10 years' worth of Medicare data, 87% of patients had just one qualifying ED index visit, independent of the 360-day moratorium on replacement. When using the prohibition against replacing patients within the 360 days after each ED index visit, 93% of patients had just one qualifying ED index visit. We were not able to conduct this analysis on 3 years' worth of data in time for this submission, but the proportion of patients with second visits in the 360 days after the ED index visit during any 3-year period is therefore likely to be roughly 6%.
- **Issue 10:** One reviewer sought some additional documentation and discussion on the 30 day and 90-360 day windows. They requested a graph that goes to the 50th week post-discharge to better understand baseline rates and variation across hospitals.
    - **Response:** We have provided some description above of the rationale behind 30-day and 90-360 day windows in response to a prior question about observed minus expected risk difference methods. We elaborate further here. The choice of these dates has to do with the natural history of major stroke after minor stroke and transient ischemic attack (TIA), which is shown below in Figure 2A (and matches the curve of post-ED-discharge stroke hospitalizations shown in Figure 2B). From an emergency department (ED) perspective, 30 days is often considered the upper limit for a time

window to include "ED-related events." Figures 2A and 2B illustrate that 30 days is a reasonable limit for attributable stroke hospitalizations, but the actual incidence rate doesn't fully stabilize until about 90 days. Thus we have eliminated the transitional risk period (30-90 days) during which the incidence rate is stabilizing out to a flat baseline.

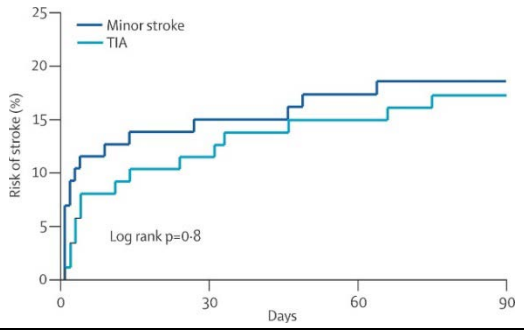


Figure 2A. Cumulative incidence curve for natural history of major stroke following TIA or minor stroke. Data are from the population-based Oxford Vascular Study as represented in Rothwell & Johnston 2006, Lancet Neurol.

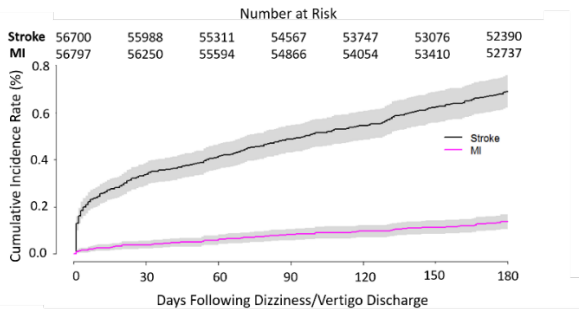


Figure 2B. Cumulative incidence of stroke hospitalizations post ambulatory (ED or other) treat-and-release as "benign dizziness." Heart attack (MI) returns are shown as a "control" comparator.

- In regards to the request to share a graph that goes to 360 days post-discharge to better understand baseline stroke risk, our team stopped, several years ago, making graphs out to 360 days because the stroke risk becomes low and linear very rapidly, and the stretch from 180 days to 360 days is not every informative. To complement the 180-day graph that we shared in section 2.b3.2 of the initial application, we did identify two earlier studies in the published literature that have looked at stroke risk for at least 1 year after discharge (Figures 3 and 4). As both figures show, the cumulative risk of a stroke turns quite linear 60 days (2 months) post-discharge.

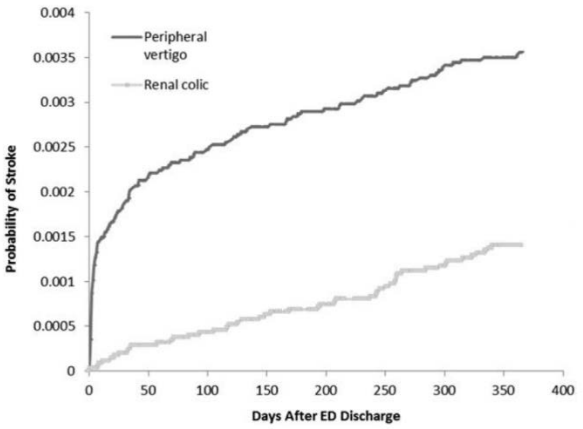


FIGURE 2: Cumulative incidence curve of stroke risk in matched patients with a discharge diagnosis of a peripheral vestibular disorder versus renal colic. ED = emergency department.

Figure 3. From Atzema et al, 2015, *Ann Neurol*.

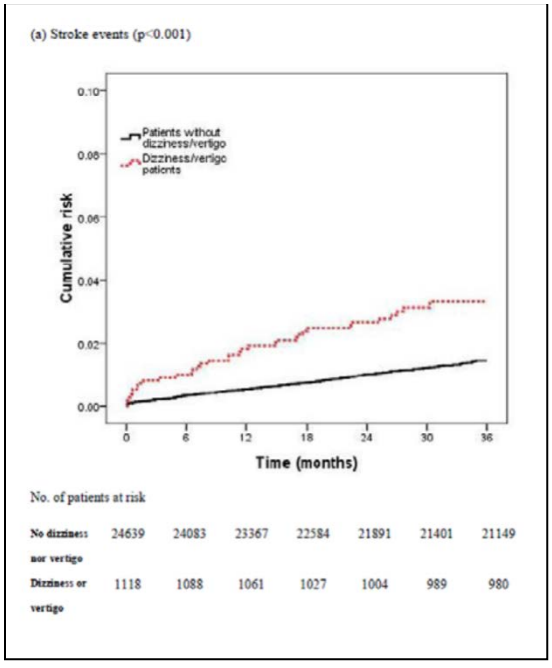


Figure 4. From Lee et al, 2012.

Measure Number: 3623

Measure Title: Elective Primary Hip Arthroplasty

Measure Developer/Steward: Acumen LLC / CMS

Validity

- **Issue 1:** Some panelists asked for clarification on the attribution methodology, specifically how episodes are attributed to main and assistant clinicians, and how the TIN-NPI attribution relates to the collaborative nature of care throughout the episode.
    - **Developer Response 1:** An episode will be attributed to both main and assistant clinicians if they billed a trigger code for the hip arthroplasty procedure. This attribution methodology ensures that the measure focuses on the role of the clinician performing the surgical procedure in line with the measure intent.
    - Clinicians can choose participate in MIPS as an individual TIN-NPI or as part of a group (TIN). Under both participation options, the measure is constructed to encourage care coordination throughout the patient's care trajectory by including the costs of clinically related services. In the SMP member's example, the attributed clinician(s) would be incentivized to coordinate with their colleagues in the academic medical center on the patient's care whether participating in MIPS as an individual TIN-NPI or as part of the attributed TIN. Based on 2018 Quality Payment Program (QPP) data released in late 2020, only 6% of eligible clinicians participated as individuals, so the TIN attribution methodology is much more widely used.[1]
- **Issue 2:** Panelists had questions about services, specifically how clinically related services are determined and how costs unrelated to the surgery/recovery are accounted for.
    - **Developer Response 2:** The measure accounts for unrelated costs by not including them in the set of assigned services; that is, the measure uses a set of detailed service assignment rules to ensure that only services that are clinically related to the role of the clinician performing this procedure are included. This includes services where the attributed clinician can influence the frequency or severity of the service.
    - The methodology for determining clinically related services involved extensive expert input in an iterative process, empirical analyses, field testing, and incorporating the perspective of persons with lived experience of the procedure. For example:
        - The Clinician Expert Workgroup members reviewed analyses of the utilization and timing of all Medicare Parts A and B services relative to the episode trigger to identify services for inclusion. To ensure clinical relatedness, they could also apply additional rules to the service, such as requiring a particular diagnosis to accompany the service.
        - Eight patient representatives provided input through structured interviews based on their experience of undergoing a hip arthroplasty. This perspective is reflected by including the costs of care services that patients experienced, including imaging and testing prior to the surgery, types of complications that patients were told by their care team could occur (e.g., infections), post-operative recovery through PAC and therapy, and routine follow-up care with the orthopedic surgeon.
    - The list of clinically related services are detailed in the Measure Codes List file on the

"Service_Assignment" tab (information on accessing the file is provided in Section S.1 of the Intent to Submit. It is available for download from the QPP Resource Library[2]).

- **Issue 3:** A panelist questioned how the measure accounts for higher frequency/severity of clinically related services if they relate to a pre-existing chronic condition.
    - **Developer Response 3**: The measure is risk adjusted to account for the different levels of patient complexity that affect resource use. The risk adjustment model includes conditions defined through 79 Hierarchical Condition Categories (HCCs), disease interactions, and a range of patient factors that clinical experts identified as being important to account for in assessing the cost of care for hip arthroplasty (e.g., osteoporosis).
- **Issue 4:** Some panelists asked for clarification of the rationale and methodology for correlating the cost measure with the MSPB Hospital measure.
    - **Developer Response 4:** NQF #2158 MSPB Hospital provides a strong conceptual relationship with the Elective Primary Hip Arthroplasty measure, given the focus on resource use across both measures and the substantial role of hospitals in hip arthroplasty procedures (e.g., in coordination of care during relevant hospital stays). It has been NQF-endorsed and used in hospital performance programs since 2013, so the measure provides a suitable external source of resource use information to validate the resource use capture of this Elective Primary Hip Arthroplasty measure. The hospital-level measure also provides a greater sample size than clinician-level measures, which are limited due to the voluntary reporting of MIPS quality measures.
    - Hospitals were associated with a clinician (TIN, TIN-NPI) if the clinician had episodes with inpatient services performed at the hospital. Of these hospitals, clinicians were then assigned the MSPB Hospital score of the hospital where the majority/plurality of their inpatient episodes/services occurred.
- **Issue 5:** A panelist questioned the exclusion for hip arthroplasties due to fracture or trauma, given the similarity in costs.
    - **Developer Response 5:** Since fracture or trauma would not be elective procedures, these are outside the scope of the measure intent. The clinical rationale for focusing on elective procedures is because these patients may be higher-risk or otherwise require additional care compared to the general patient population. This is shown in the testing analyses provided in the Testing Form Appendix, Table 2b2.2. To clarify, these episodes have a substantially higher observed cost (by almost a third) and observed to expected cost ratio than final episodes (1.35 compared to 0.98 and 0.97 for TINs and TIN-NPIs, respectively).

References

1. CMS, "2018 Quality Payment Program Experience Report," Quality Payment Program (September 2020), https://qpp-cm-prod-content.s3.amazonaws.com/uploads/1091/2018%20QPP%20Experience%20Report.pdf
2. CMS, "Resource Library," Quality Payment Program, https://qpp.cms.gov/resources/resource-library. Direct download of Measure Codes List files: https://qpp-cm-prod-content.s3.amazonaws.com/uploads/1262/2021-cost-measure-codes-lists.zip

Measure Number: 3625

Measure Title: Non-Emergent Coronary Artery Bypass Graft (CABG)

Measure Developer/Steward: Acumen LLC / CMS

Validity

- **Issue 1:** Panelists asked about service assignment, specifically the list of assigned services, how clinical relatedness was determined, and how other related services could be added.
    - **Developer Response 1:** The list of clinically related services are detailed in the Measure Codes List file on the "Service_Assignment" tab (link is provided in Section S.1 of the Intent to Submit form. It is available for download from the QPP Resource Library[1]).
    - The methodology for determining clinically related services involved extensive expert input in an iterative process, empirical analyses, field testing, and incorporating the perspective of persons with lived experience of the procedure. For example:
        - The Clinician Expert Workgroup members reviewed analyses of the utilization and timing of all Medicare Parts A and B services relative to the episode trigger to identify services for inclusion. They could also use additional rules to ensure clinical relatedness, such as requiring a diagnosis on the service.
        - Five patient representatives provided input through structured interviews based on their experience of undergoing a CABG. This perspective is reflected by including the care services that patients experienced, including imaging, testing, wound care (e.g., to prevent infections), cardiac rehabilitation through different types of PAC, and follow-up visits with the surgeon.
    - The service assignment rules are revisited regularly as part of measure maintenance to ensure the codes for assigned services are up-to-date and remain clinically relevant; additional related services can be added through this process (e.g., certain telehealth services were added in 2020).
- **Issue 2:** Panelists asked for clarification of the measure scope and trigger methodology to exclude emergent CABG procedures.
    - **Developer Response 2:** The measure scope focuses only on non-emergent CABG procedures, based on expert clinical input and empirical analyses. The Cardiovascular Disease Management Clinical Subcommittee and CABG Clinician Expert Workgroup identified that focusing on patients without urgent indications ensures a more clinically homogenous patient cohort. Reducing heterogeneity can help improve the validity of the cost measure by removing sources of variation outside clinician influence and can prevent unintended consequences of measuring clinician cost performance when treating unique patient populations. This clinical rationale is corroborated by the results of exclusions analyses, where emergent CABG episodes have both a higher mean observed cost and a wider range of observed to expected cost ratios than the final episodes after non-emergent CABG trigger logic and exclusions are applied. Further data can be found in Section 2b2.2 of the Measure Testing Form.
    - The trigger logic and exclusions ensure that appropriate procedures are captured in the measure given the measure intent. Since isolated CABG surgery and CABG surgery with

concurrent aortic valve replacement take place during an inpatient stay, the trigger logic requires the presence of either a cardiac valve MS-DRG (i.e., MS-DRGs 216-221) or a coronary bypass MS-DRG (i.e., MS-DRGs 231-236) to ensure that the reason for admission was the CABG surgery. The trigger methodology also has a CPT/HCPCS code exclusion which removes episodes where other cardiac-related procedure occurs at the same time as the initial trigger procedure for heart artery bypass surgery, as these cases may be more complex, may indicate emergent situations, or may complicate the evaluation of care for the CABG procedure alone.

- **Issue 3:** A panelist asked for the rationale for including risk adjustors associated with homebound status.
    - **Developer Response 3: Workgroup members recommended frailty-related risk** adjustors based on their clinical expertise and empirical analyses. As mentioned in Section 2b3.1.1 and 2b3.3a in the Measure Testing Form, the workgroup recommended accounting for patient cohorts, like those with dementia, walking aids, or home health services as these pre-existing factors affect the cost of care for CABG. For example, patients with dementia are more likely to have increased costs following the procedure due to complications and additional required services, including post-acute care and rehabilitation. Patients receiving home health services or who have walking aids are indicators of frailty, which are associated with higher resource utilization after CABG surgery.
    - The Clinician Expert Workgroup also believed that these and other custom risk adjustors were important for clinical face validity, as shown through the face validity vote where all 9 members agreed with the risk adjustors (7: strongly agree, 2: moderately agree).

References

1. CMS, “Resource Library,” Quality Payment Program, https://qpp.cms.gov/resources/resource-library. Direct download of Measure Codes List files: https://qpp-cm-prod-content.s3.amazonaws.com/uploads/1262/2021-cost-measure-codes-lists.zip

Measure Number: 3626

Measure Title: Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels

Measure Developer/Steward: Acumen LLC / CMS

Reliability

- **Issue 1**: One panelist questioned the use of a 10-episode testing volume threshold and whether this is consistent with specifications.
    - **Developer Response 1:** Testing results are presented for TINs and TIN-NPIs with at least 10 episodes; we use this threshold as this is the case minimum as used for the measure in MIPS. Sections 1.5, 1.6, 2a2.2, and 2b2.2 of the Measure Testing Form note the use of the 10-episode case minimum.

Validity

- **Issue 2:** Panelists asked for clarification on the attribution methodology, specifically as it relates to the role of members of the care team (e.g., main surgeon, co-surgeon, anesthesiologist).
    - **Developer Response 2:** The attribution methodology focuses on the clinician(s) performing the lumbar spine fusion procedure by attributing an episode to the clinician(s) who bill the trigger code (CPT/HCPCS procedure code). This can be both the main and assistant clinician. We use this methodology as the measure intent is to assess costs related to the role of the clinician performing the surgical procedure. Since the role of an anesthesiologist or CRNA is distinct from performing the surgery itself, this measure does not attribute episodes to members of the care team who do not bill the trigger procedure.
- **Issue 3:** One panelist asked about the role of sub-groups and testing results at the sub-group level.
    - **Developer Response 3:** The use of sub-groups is to improve the clinical comparability of the cost measure: the patient cohort is stratified into three mutually exclusive and exhaustive sub-groups based on the type of procedure. Specifically, the sub-groups are determined by how many segments of the spine are fused together, or what "level" the procedure is (e.g., a one-level procedure fuses one segment to join two vertebrae). The Intent to Submit Form, S.7.2 in describes each sub-group in further detail in Step 1. The risk adjustment model is run separately within each sub-group to account for differences in resource use that may stem from the complexity of the procedure. This means that patients are only compared with other patients undergoing the same type of surgery. The final measure score combines results from all sub-groups to evaluate overall performance across procedure types: please see Appendix B of the Measure Information Form (MIF) for an example illustrating the measure calculation (information on accessing the specifications is provided in Section S.1 of the Intent to Submit. It is also available for download from the QPP Resource Library).[1]
    - The Testing Form Appendix includes results of risk adjustment testing at the sub-group level: please see Tables 2b3.1.1 and 2b3.4b.
- **Issue 4:** A panelist asked whether Skilled Nursing Facility (SNF) costs are included. Panelists also asked for documentation of the clinically related services.
    - **Developer Response 4:** The measure includes SNF costs where the SNF claim's qualifying inpatient stay is the same as the trigger inpatient procedure. This ensures that SNF is only assigned to an episode where it is closely related to the inpatient surgical procedure. This is detailed in the MIF, Section A.3.
    - The list of clinically related services are detailed in the Measure Codes List file on the "Service_Assignment" tab (information on accessing the file is provided in Section S.1 of the Intent to Submit. It is available for download from the QPP Resource Library).[2]
- **Issue 5:** A panelist asked for the number of patients excluded from the measure under each type of exclusion.
    - **Developer Response 5:** Please see Table 2b2.2 in the Testing Form Appendix for the number of episodes and beneficiaries with the exclusion characteristic. Please note that

this table lists different types of exclusions for: data cleaning (rows from ISO 2 to ISO 8), patients who are not part of the measure intent or are not able to be fairly compared with the overall patient cohort (rows from ISO 9 to ISO 17), and exclusions due to the case minimum or statistical outliers (rows from ISO 18 to ISO 20).

References

1. CMS, "Resource Library," Quality Payment Program, https://qpp.cms.gov/resources/resource-library. Direct download of MIFs: https://qpp-cm-prod-content.s3.amazonaws.com/uploads/1261/2021-cost-measure-information-forms.zip
2. CMS, "Resource Library," Quality Payment Program, https://qpp.cms.gov/resources/resource-library. Direct download of Measure Codes List files: https://qpp-cm-prod-content.s3.amazonaws.com/uploads/1262/2021-cost-measure-codes-lists.zip

Subgroup 3

*M*easure Number: 0500

Measure Title: Severe Sepsis and Septic Shock: Management Bundle

Measure Developer/Steward: Henry Ford Hospital

Reliability

- **Issue 1:** Sampling. One reviewer asked about the method used for population sampling.
    - Developer Response 1: The SEP-1 measure specifications refer to the "Population and Sampling Specifications" section of the Hospital Inpatient Reporting Program specifications manual, which provides more detailed guidelines on sampling approaches.
- **Issue 2:** Measure updates. One reviewer expressed concern that the most recent testing report includes two quarters of data and that changes in measure specifications may affect reliability over time.
    - **Developer Response 2:** The SEP-1 measure, like all measures, is regularly updated to incorporate the most recent relevant evidence, address clinician and abstractor feedback, clarify abstraction guidance, reduce abstractor burden, and capture changes in codes. The measure stewards carefully consider the potential impact of measure updates that would affect measure stability. Since 2018, the measure changes have focused on clarifying the abstraction guidance, with only minor changes to the measure logic, numerator, and denominator. The processes used to propose and approve changes for SEP-1 include an analysis of test case scenarios and available data to confirm the changes do not impact measure results, which is consistent with update processes for other measures in CMS programs. The measure stewards and developers consults with an Expert Work Group to discuss the effects of proposed changes to the measure on provider burden and implications for patient care. We used data from Q3 2018 and Q4 2018 as a sample from the most recently available data at the time testing occurred. Testing results from the 2017 submission (which covers Q4 2015 – Q2 2016) and the 2021 submission demonstrate acceptable reliability. We also note that the

measure is used in a CMS pay for reporting program, so performance on SEP-1 scores **do not** affect hospitals' financial performance.

- **Issue 3:** Severe Sepsis Presentation Time. One reviewer asked how "Time 0" is determined for severe sepsis.
    - **Developer Response 3:** The Severe Sepsis Presentation Time data element in the Data Dictionary includes detailed guidance on how this time is determined. Severe Sepsis Presentation Time is determined by the earliest time that the final clinical criterion for severe sepsis (as defined by the Severe Sepsis Present data element) was noted or the earliest time the physician/APN/PA documented severe sepsis. The Severe Sepsis Present data element specifies that three clinical criteria, which include documentation of an infection, two or more Systemic Inflammatory Response Syndrome (SIRS) criteria, and a sign of organ dysfunction must be met within six hours of each other.
- **Issue 4:** Patient characteristics. One reviewer asked for patient-level descriptive information for data used in the reliability testing in section 1.6.
    - **Developer Response 4:** Table 1.6.1 of the 2021 testing attachment contains patient-level descriptive information for patients included in the reliability analysis.
- **Issue 5:** Data element reliability. Two reviewers discussed the value of data element reliability analyses in their responses.
    - **Developer Response 5:** The testing attachment instructions indicate that reliability must be demonstrated for the composite performance measure score for composite measures. The testing attachment included results for data element validity, which compares hospital-abstracted scores with CDAC data, which serves as the "gold standard." According to the NQF guidelines, if data element validity is conducted, data element reliability is not required.
- **Issue 6:** Distribution of reliability scores. One reviewer asked for the distribution of reliability scores.
    - **Developer Response 6:** Section 2a2.3 of the 2021 Submission shows the distribution, including the 25th percentile, median, and 75th percentile of signal to noise ratios by quarter.
- **Issue 7:** Split-sample analysis. One reviewer recommended including a split-sample or stability of classification analysis.
    - **Developer Response 7:** In section 2a2.3 of the 2021 submission, the analysis found that there was not a consistent trend in mean reliability by hospital denominator size, but that the reliability score for each decile was over 0.70.

Validity

- **Issue 1:** Measure score validity analysis. One reviewer discussed questions about the measure score validity analysis.
    - **Developer Response 1:** One reviewer expressed concerns about the association between pass rate and mortality rate. Table 2b1.3.11 of the 2021 submission includes a patient-level analysis indicates a significant association between SEP-1 compliance and mortality (OR = 0.77, 95% CI = 0.76-0.79). The graphs displaying mortality by pass rates and mortality rates show an inverse relationship overall. Table 2b1.3.8 from the 2021

submission indicates higher mortality rates for hospitals with lower performance. The reviewer noted an error in definition of the p-value in the methods section of the 2017 submission, but this error was corrected in the 2021 submission, and the p-value was calculated correctly in both submissions. The reviewer also describes that the two proportion z-test assessing significant differences in pass rate is guaranteed to be significant; however, that is not necessarily true as the hypothesis test is assessing whether there is a statistically significant difference between deciles.

- **Issue 2:** Hospital-level vs. patient level analyses. One reviewer noted concerns of aggregation bias for the measure-score level analysis.
    - **Developer Response 2:** The validity analyses describe relationship between performance and mortality both at the hospital and patient level. The patient-level propensity score analysis in section 2b1.3.11 supports that patients who receive SEP-1 compliant care have lower odds of mortality compared to patients who do not.
- **Issue 3:** Chance-corrected agreement statistics. One reviewer asks for chance-corrected agreement statistics to be added to the submission.
    - **Developer Response 3:** Section 2b1.3 of the 2021 submission includes kappa values to assess chance-adjusted agreement.
- **Issue 4: Risk adjustment.** Two reviewers asked for justification for the risk adjustment approach and one noted that the developer should explore whether demographic factors such as race and gender have an influence on the performance score.
    - **Developer Response 4:** Section 2b4.2 of the 2021 submission assesses disparities based on race, gender, payer, and age. To our knowledge, there is no evidence supporting risk adjustment of this process measure based on these characteristics.
- **Issue 5:** Missing data. One reviewer asked about analysis of cases that were not included in the abstraction tool.
    - **Developer Response 5:** The measure calculation only includes cases which are entered into the abstraction tool and submitted to the CMS Clinical Data Warehouse, so the measure stewards and measure developers do not have access to cases which were rejected from the measure algorithm due to missing data.
- **Issue 6:** Coding. One reviewer asked whether there is an analysis assessing the relationship between clinician documentation, coding, and care provision.
    - **Developer Response 6:** Data collected for this measure is based upon information documented in the patient's medical record. Other data related to accuracy of documentation and coding is not available.
- **Issue 7:** Face validity. One reviewer noted that "application of the measure" may be where threats to face validity occur."
    - **Developer Response 7:** To our knowledge, there is not evidence indicating threats to face validity of the measure in the application of the measure.
- **Issue 8:** Data element validity. Some reviewers commented on data element validity for time variables.
    - **Developer Response 8:** Table 2b1.3.2 of the 2021 submission shows the correlation between the hospital-abstracted and CDAC-abstracted times, which serves as the "gold

standard." Each of the correlations are over 0.80, indicating high correlation and supporting data element validity.

Composite

- **Issue 1: Composite analysis.** Two reviewers commented on the relationship between the components in the measure.
    - **Developer Response 1:** The processes in the measure are linked steps, and patients must receive each of the care elements that they are eligible for in order to pass the measure. The NY State Department of Health study included analysis of the impact of completing specific "processes of care" using SEP-1 data elements on outcomes such as mortality. The NQF composite measure evaluation guidance notes that "clinical justification (e.g. correlation of the individual component measures to a common outcome measure) is an acceptable method to assess the composite construction approach.

Other General Comments

- **Issue 1: Sepsis definitions.** One reviewer discussed an alternate definition of sepsis using the Sequential (Sepsis-Related) Organ Failure Assessment (SOFA).
    - **Developer Response 1:** The intent of SEP-1 is early identification and treatment. Evidence from literature suggests that SIRS based criteria, which SEP-1 uses, is better at early detection than SOFA/qSOFA, whereas SOFA/qSOFA is better at predicting cases that are at higher risk for mortality (Haydar, 2017; Waligora, 2020; Serafim 2018). Based on the available evidence, the measure stewards, SEP-1 Expert Work Group, and a measure steward and Infectious Diseases Society of America (IDSA) collaborative work group all agree SEP-1 should continue to use SIRS based criteria. Additionally, there has not been wide spread acceptance of SOFA/qSOFA. SEP-1 uses a two-step approach to identifying patients who are eligible for the denominator. The initial patient population is based upon ICD-10 codes for sepsis, severe sepsis without shock, and severe sepsis with shock. Abstractors then confirm whether to include each patient case based on meeting the SIRS based criteria in the Severe Sepsis Present data element. Cases are not included in the measure unless they meet both the ICD-10 code requirements and the SEP-1 definition and criteria. This is consistent with the intent of early identification of cases to determine whether evidenced based appropriate care was provided. The measure emphasizes the importance of early clinical identification of sepsis, severe sepsis and septic shock regardless of the method used by the individual clinician and hospital system. The fact that SEP-1 has not changed the definition or the way severe sepsis or septic is identified for purposes of the measure means the measure continues to have a stable population, which would not impact reliability and validity.

References:

1. Haydar, S., Spanier, M., Weems, P., Wood, S., & Strout, T. (2017). Comparison of QSOFA score and SIRS criteria as screening mechanisms for emergency department sepsis. *The American Journal of Emergency Medicine, 35*(11), 1730-1733.

2. Waligora, G., Gaddis, G., Church, A., & Mills, L. (2020). Rapid Systematic Review: The Appropriate Use of Quick Sequential Organ Failure Assessment (qSOFA) in the Emergency Department. *The Journal of Emergency Medicine, 59*(6), 977-983.

3. Serafim, R., Gomes, J. A., Salluh, J., & Póvoa, P. (2018). A comparison of the quick-SOFA and systemic inflammatory response syndrome criteria for the diagnosis of sepsis and prediction of mortality: a systematic review and meta-analysis. *Chest, 153*(3), 646-655.

Measure Number: 0674

Measure Title: Percent of Residents Experiencing One or More Falls with Major Injury (Long Stay)

Measure Developer/Steward: The Centers for Medicare & Medicaid Services (CMS)

Reliability

- **Issue 1:** The 2008 RAND testing is old. One panel member asked how can we be sure that the data element reliability and validity have been maintained?
    - **Developer Response 1:** The MDS 3.0 item set has remained stable since RAND created the recommended MDS 3.0 form in 2008, with the exception of select changes in item specifications and the addition of some new items that do not affect this measure. In particular, the Falls with Major Injury item has the same look-back period and the same item wording in the latest MDS 3.0 form and the 2008 recommended form.
    - The authors of the RAND study also conducted an evaluation of the MDS 3.0 form in 2012 to determine whether their revisions improved reliability, validity, resident input, and clinical utility, all while decreasing collection burden. The results demonstrated that the reliability for research nurse to research nurse comparisons and for research nurse to facility staff comparisons was good or excellent for most MDS items including falls with major injury, and there was increased validity compared with MDS 2.0.[1]
- **Issue 2:** One panel member mentioned that comparing community nurses to gold standard nurses seems like a test of validity rather than reliability.
    - **Developer Response 2:** The RAND study uses inter-rater agreement to assess the reliability and reproducibility of MDS items (as suggested in standard reliability studies, such as Landis and Koch (1977)[2]). The extent of inter-rater agreement across nurses reflects the reliability of a data element in terms of consistently or precisely reflecting a patient's circumstances. We agree that the fact that this comparison examines community nurses and "gold standard" nurses also speaks to validity of the data elements, in terms of accurately reflecting a patient's circumstances. In addition to the high level of agreement observed in the RAND study, 88% of nurses who participated in this MDS national study said that the fall-related injury definitions were clear. 94% felt that facilities' falls documentation should include the information needed to complete the section accurately.

Validity

- **Issue 1:** One panel member asked if the variation by state analyses were done with 3-level hierarchical (patient/facility/state) models or an ANOVA on the facility level scores.
    - **Developer Response 1:** The variation by state analyses were done with an ANOVA on the facility level scores.

Other General Comments

- **Issue 1:** One panel member recommended a volume threshold for the measure reporting.

- **Developer Response 1:** There is an existing measure reporting threshold of 20 stays for this measure. The measure is publicly reported on the Care Compare site, and the quality measure data are available to download at data.cms.gov. MDS quality measures reported on data.cms.gov contain quality measure scores for specific nursing homes, including the 4-quarter score average and scores for each individual quarter, while Care Compare displays the 4-quarter score average only. The 4-quarter score average is used for CMS's Five-Star Rating program, and the Five-Star program requires the measure denominator to include at least 20 residents' assessments across four quarters of data. The intent of this requirement is to enhance measurement stability and reliability beyond a one-quarter measure.

References

1. Saliba, D., & Buchanan, J. (2012). "Making the Investment Count: Revision of the Minimum Data Set for Nursing Homes, MDS 3.0." Journal of American Medical Directors Association 13(7): 602-10. https://doi.org/10.1016/j.jamda.2012.06.002.
2. Landis, J R, and G G Koch. "The measurement of observer agreement for categorical data." Biometrics vol. 33,1 (1977): 159-74.

Measure Number: 0679

Measure Title: Percent of High Risk Residents with Pressure Ulcers (Long Stay)

Measure Developer/Steward: CMS

Reliability

- **Issue 1:** There are some concerns that the 2008 RAND study used for data element testing is old.
  - **Developer Response 1:** The MDS 3.0 item set has remained stable since RAND created the recommended MDS 3.0 form in 2008, with the exception of select changes in item specifications and the addition of some new items that do not affect this measure. In particular, the Pressure Ulcer item has the same item wording in the latest MDS 3.0 form and the 2008 recommended form.
  - The authors of the RAND study also conducted an evaluation of the MDS 3.0 form in 2012 to determine whether their revisions improved reliability, validity, resident input, and clinical utility, all while decreasing collection burden. The results demonstrated that the reliability for research nurse to research nurse comparison and for research nurse to facility staff comparison was good or excellent for most MDS items including pressure ulcer, and there was increased validity compared with MDS 2.0.[1]
- **Issue 2:** One panel member mentioned that comparing community nurses to gold standard nurses seems like a test of validity rather than reliability.
  - **Developer Response 2:** The RAND study uses inter-rater agreement to assess the reliability and reproducibility of MDS items (as suggested in standard reliability studies, such as Landis and Koch (1977)[2]). The extent of inter-rater agreement across nurses reflects the reliability of a data element in terms of *consistently* or *precisely* reflecting a patient's circumstances. We agree that the fact that this comparison examines community nurses and "gold standard"

nurses also speaks to validity of the data elements, in terms of *accurately* reflecting a patient's circumstances. In addition to the high level of agreement observed in the RAND study, 89% of nurses who participated in this MDS national study said that the pressure ulcer definitions were clear. 83% felt that the form was easy to use for reporting pressure ulcers at different stages.

Validity

- **Issue 1:** One panel member asked if the variation by state analyses were done with 3-level hierarchical (patient/facility/state) models or an ANOVA on the facility level scores.
    - **Developer Response 1:** The variation by state analyses were done with an ANOVA on the facility level scores.

Other General Comments

- **Issue 1:** One panel member recommended a volume threshold for the measure reporting.
    - **Developer Response 1:** There is an existing measure reporting threshold of 20 stays for this measure. The measure is publicly reported on the Care Compare site, and the quality measure data are available to download at data.cms.gov. MDS quality measures reported on data.cms.gov contain quality measure scores for specific nursing homes, including the 4-quarter score average and scores for each individual quarter, while Care Compare displays the 4-quarter score average only. The 4-quarter score average is used for CMS's Five-Star Rating program, and the Five-Star program requires the measure denominator to include at least 20 residents' assessments across four quarters of data. The intent of this requirement is to enhance measurement stability and reliability beyond a one-quarter measure.

References

1. Saliba, D., & Buchanan, J. (2012). "Making the Investment Count: Revision of the Minimum Data Set for Nursing Homes, MDS 3.0." *Journal of American Medical Directors Association* 13(7): 602-10. https://doi.org/10.1016/j.jamda.2012.06.002.
2. Landis, J R, and G G Koch. "The measurement of observer agreement for categorical data." Biometrics vol. 33,1 (1977): 159-74.

Measure Number: 2902

Measure Title: Contraceptive Care - Postpartum

Measure Developer/Steward: HHS Office of Population Affairs

Reliability

- **Issue 1:** The following attachment was not provided: NQF_2902_Codes_2021.xlsx
    - **Developer Response 1:** The codes were submitted along with the testing attachment.
- **Issue 2:** It is not clear what the nature of the hierarchical structure of the primary and sub-measure is. What impact on the score does this structure have? No justification was provided for excluding deliveries that did not end in a live birth (i.e., miscarriage, ectopic, stillbirth or induced abortion). Aren't these women also at risk of having an unintended pregnancies with its associated poor outcomes?

    - **Developer Response 2:** The primary measure numerator includes all methods of contraception that require a prescription and sterilization, while the sub-measure numerator focuses on the provision of a subset of prescription contraceptives, LARC methods. The LARC sub-measure is calculated separately from the most and moderately effective methods because the interpretation is different. The LARC measure is a floor measure, which we define as a measure that identifies instances where LARC is not provided at all or at very low rates (i.e., less than 2%).
    - Non-live births were excluded because this measure is trying to capture the population receiving postpartum care. Postpartum care service is usually provided by a specific group of providers to women who have a live birth and women whose pregnancies do not end in a live birth usually do not receive this service. Women who needed contraceptive care after a non-live birth are captured by the other two contraceptive measures, 2903 (Contraceptive Care – Most & Moderately Effective Methods) and 2904 (Contraceptive Care – Access to LARC).
- **Issue 3:** A detailed description of the method used for reliability testing was supposed to be demonstrated in the Appendix, but I could not locate it within the submission materials.
    - **Developer Response 3:** The method was described in detail in Appendix C.
- **Issue 4:** Given that the minimum sample size of 75 patients yields sufficient reliability, it may be difficult to achieve for some provider groups.
    - **Developer Response 4:** We agree that the threshold required to reach sufficient reliability will exclude some small provider groups. However, we do not think this will have a large impact on the measure usage because most of the groups serving less than 75 patients in a year are individual providers who registered as a group practice, and the measures are not specified to be calculated at the individual provider level.
- **Issue 5:** The level of analyses specified (Clinician: Group/Practice, Health Plan, Population: Regional and State) does not match the levels reported in the testing form (group/practice, health plan, public health region). Please clarify this and for which of these levels was the measure specified.
    - **Developer Response 5:** The measure was specified at the health plan and public health region levels, as tested in the original application. In the current re-endorsement application, we tested the measure at these two levels as well as an additional group/practice level.

Validity

- **Issue 1:** Did not provide any empirical validity testing for health plans & population.
    - **Developer Response 1:** We did not conduct score level validity testing at the health plan or region levels due to the limited numbers of units (n≤21) at these levels that are not sufficient for correlation testing.
- **Issue 2:** Sensitivity for contraceptive patch (numerator variable) and live birth in the last 2 months (exclusion variable) is quite low. This is concerning as it will affect identification of both outcome and cohort.
    - **Developer Response 2:** We did not conduct data element level validity for this measure. This comment seems to apply to another contraceptive measure (2903). We address

this concern in the response for 2903.

- **Issue 3:** The measure does not account for patient preference or Academy of Breastfeeding statement. Some regions/ patients groups/patients may not view these intermediate outcomes as favorable - infringement on patient autonomy.
    - **Developer Response 3:** Although some postpartum care providers may be reluctant to offer the most effective hormonal methods because of the 2015 Academy of Breastfeeding Medicine's (ABM) statement on contraceptive choice during breastfeeding, ABM did endorse ACOG's 2018 Committee Opinion No. 736 and its recommendations on contraception during breastfeeding. Endorsed also by the American College of Nurse-Midwives, the National Association of Nurse Practitioners in Women's Health, the Society for Academic Specialists in General Obstetrics and Gynecology, and the Society for Maternal–Fetal Medicine, this current clinical guideline states that providers should "review theoretical concerns regarding hormonal contraception and breastfeeding, within the context of each woman's desire to breastfeed and her risk of unplanned pregnancy". This recommendation for patient-centered care allows postpartum women to make an informed, autonomous decision which reflects their needs and preferences for both breastfeeding and contraception. ACOG Committee Opinion No. 756 (also released in 2018) echoes this guideline.
    - OPA acknowledges the concerns about accounting for patient autonomy with this measure. OPA strongly believes that the goal of providing contraception should never be to recommend any one method or class of methods over women's individual choices. Thus, the contraceptive care measures are designed to encourage providers to offer those clients seeking contraception the full range of methods. Like #2903 and #2904, #2902 is specified for use in administrative claims data, which has a limitation in that it does not contain information about patient preferences. Thus, OPA currently works to indirectly account for client choice and to strongly promote patient-centered contraceptive care through the Providing Quality Family Planning (QFP) guidelines [Gavin & CDC 2014]. Jointly published by OPA and CDC, QFP states that it is important that these contraceptive services are provided in a client-centered manner that treats each person as a unique individual with respect, empathy, and understanding, providing accurate, easy-to-understand information based on the client's self-identified needs, goals, preferences, and values. Clinics can utilize these guidelines with the contraceptive measures to deliver patient-centered contraceptive care.
    - The NQF #2902 primary measure is designed so that all methods of contraception that require a prescription and sterilization are included in the numerator, which are treated as being of equal value during measure calculation. Hence, the numerator represents a wide range of methods from which clients can choose.
    - To encourage providers to deliver family planning care in a fully person-centered, non-coercive manner, OPA also states on its public website (https://opa.hhs.gov/research-evaluation/title-x-services-research/contraceptive-care-measures) that the most and moderately effective methods measures should not be used to encourage high utilization rates. It also emphasizes that the LARC sub-measure would be an

inappropriate measure to implement in a pay-for-performance context because it is a floor measure. The goal of the NQF #2902 sub-measure is to ensure access to LARC methods in the postpartum period by monitoring very low rates of provision (i.e., below 2%).

- Research indicates that patients receiving client-centered care may feel motivated to continue seeking reproductive health care for contraception and if they become pregnant, prenatal care and birth [Gomez 2017]. To broadly promote patient-centered reproductive health services, OPA delivers online education on this topic via a grant to the Reproductive Health National Training Center website (https://rhntc.org/resources/contraceptive-counseling-and-education-elearning). Recently updated in 2020, this on-demand training module is available to all providers.
- After NQF endorsed the contraceptive provision measures, OPA demonstrated its commitment to patient-centered contraceptive care by providing funding to the University of California San Francisco (UCSF) to develop a patient-reported outcome performance measure (PRO-PM) assessing the degree to which patient needs, values, and preferences are prioritized in the counseling encounter. After the initial year of funding, UCSF secured private funding to continue the project. Recently endorsed by NQF in December 2020 as the Person-Centered Contraceptive Counseling (PCCC) measure, this measure facilitates proper interpretation of the contraceptive provision measures by allowing organizations to observe variations in patient experience that occur with changes in provision of most or moderately effective contraception, including LARC methods. Health care providers can then ensure that increases in provision are not associated with a negative patient experience; ideally, improved provision would be linked to better patient experience. Due to the distinct structure of client-centered contraceptive counseling in pregnant individuals (i.e., counseling occurs over multiple visits prior to delivery), the PCCC is a visit-specific measure that focuses on non-pregnant clients. It has not been comprehensively tested in pregnant patients. Currently conducting research to operationalize the 'tandem use' of the new PCCC measure with developmental electronic clinical quality measure (eCQM) versions of NQF #2903 and #2904, UCSF plans to investigate the possibility of utilizing a contraceptive care PRO-PM with a developmental eCQM version of NQF #2902.

- **Issue 4:** The main concern I have is with the exclusion of live birth deliveries that occurred during the last 2 months of the measurement year, which contributes the large majority of excluded cases. I think there are ways to avoid this exclusion as noted above, by modifying the included months for a given year to include a full year. Additionally, care providers might be negatively motivated to meet the required performance for the end-year births knowing that they will be excluded. No testing was conducted to assess the impact of this exclusion on the performance rate. It is therefore recommended that developers assess performance rates that include a full 12 months from a given year (Nov 1st to Oct 31st) compared to the 10 month period assessed, and add this information to the submission as complimentary material.
  - **Developer Response 4:** OPA acknowledges the concern and will investigate the issue in future submissions. OPA does not have access to two consecutive years of data in order

to provide this calculation in the current time frame.

- While unable to calculate the impact for this submission, a small percentage of women (around 16%) gave birth in the last 2 months of the year and would thus have a reduced impact on the overall yearly rates; and we would not expect care providers to consider this measure-level exclusion or to perform services differently in a patient-centered setting. Further, the measure is not specified for use at the individual provider level (and is not for use as pay-for-performance) even should the service performance decrease for a small number of providers. In addition, the intention of the measure is to provide comparison of rates across measurement years to detect changes in clinic-or-higher-level performance over time. We do not expect this comparison to be affected as any impact of this exclusion on the rates should be consistent across years. Therefore, we think the impact of this exclusion should be minimal.

- **Issue 5:** Empirical validity demonstrated poor correlation with comparative with some of the measures. Difficult to pass entire set with some results very poor. The overall measure contains multiple measures applied to different entities. Some of measure groups demonstrated poor validity. Authors caution how measures should be used but once they are endorsed, there is no control if CMS adopted them into a payment program. I have concerns of passing the entire measure when some of them performed poorly on validity.
  Validity was not strong across groups, methodologies, locations, ages and struggled between low double digits and just barely adequate double digits.
  - **Developer Response 5:** For the NQF #2902 primary measure, no benchmark is established, and we do not expect scores to reach 100%. We also emphasize that the NQF #2902 LARC sub-measure is a floor measure with scores of greater than or equal to 2% indicating some LARC access; this measure would be an inappropriate measure to implement in a pay-for-performance context. Our public-facing web site explicitly states this guidance and OPA provides the same guidance when consulted with this question.
  - During our analysis, we did note that the #2902 sub-measure for LARC provision within three days postpartum has non-significant weaker correlations with the selected comparison metrics due to a very low measure score (i.e., less than 2%). We hypothesize that this 3-day postpartum rate is lower because immediate postpartum LARC provision is a relatively new clinical practice. While supported by ACOG guidelines (Committee Opinion No. 670), this guidance was released after NQF first recommended endorsement for #2902. Thus, this sub-measure attempts to estimate access to a service that is in various stages of implementation across entities. The lower numbers of women obtaining immediate postpartum LARC within and across health care delivery systems have been explained by recent research that has highlighted the unique challenges of delivering this care in the hospital setting [Ling 2020, Moniz 2016, Moniz 2017]. These challenges currently persist even with recently adopted state Medicaid reimbursement policies. Barriers to access documented in these studies include provider attitudes and their lack of clinical training, the devices not being available onsite at the time of delivery, continued reimbursement challenges, concern about the cost of implementation activities, and uncertainty about how to provide client-centered

care in this setting. We believe this measure needs additional time to be used to ensure women have access to immediate postpartum LARC while hospitals are working to implement this clinical practice amid continued barriers.

- **Issue 6:** The justification for not risk-adjusting this measure other than age group stratification is, to my view, weak. I agree that ideally, no adjustment would be best. However, this assumes that the disparities identified between different patient groups in the use of most and moderately effective and LARC methods of contraception are within the provider's control. To justify no risk-adjustment, some level of evidence that supports this assumption should be presented. Also, no testing was provided to support the stratification approach (2b3.5. & 2b3.9). Please add this information.
- I do not understand why age-only risk stratification is offered when other demographic variables are named as potential source of variation and may be considered as risk-adjustors.
    - **Developer Response 6:** We acknowledge the concern on this issue but believe that risk adjustment isn't necessary for this measure. Although it is true that the rates differ by certain demographics characteristics, the underlying factors driving the differences are likely patient access to the service and/or patients' preference, which cannot be clearly or consistently attached to demographic characteristics. The differences in rates reflect differences in access of service by sociodemographic characteristics (financial and otherwise), and may not exist with equal access/quality of service in a patient-centered setting.
    - We calculate the measure by age group so that the measure scores for adolescents and adults for the purposes of quality improvement (QI), but not as a method of risk adjustment. Assessing the measure scores by age group is helpful because contraceptive programs focused specifically on adolescents and preventing teenage pregnancy exist. Successful QI in these programs and interventions depend on the availability of contraceptive rates among adolescents.
    - We also publish programs on our website that allow measure users to calculate rates by other demographic variables (e.g., marital status, race/ethnicity) easily.
- **Issue 7:** Even at the same measure level, the results seemed to be quite different. For example, mean rate for facilities in PPFA was 0.612 while mean rate for facilities in NYP was 0.427. It is important to establish consistent data element reliability for critical data elements across data sources before attempting to compare results based on different data sources.
    - **Developer Response 7:** This comment does not seem to apply to this measure. The rates quoted here are from another contraceptive measure (2903).

References

1. Gavin, L., Moskosky, S., Carter, M., Curtis, K., Glass, E., Godfrey, E., Marcell, A., Mautone-Smith, N., Pazol, K., Tepper, N., Zapata, L., & Centers for Disease Control and Prevention (CDC) (2014). Providing quality family planning services: Recommendations of CDC and the U.S. Office of Population Affairs. *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports*, *63*(RR-04), 1–54.

2. Gomez, A. M., & Wapman, M. (2017). Under (implicit) pressure: young Black and Latina women's perceptions of contraceptive care. Contraception, 96(4), 221–226. https://doi.org/10.1016/j.contraception.2017.07.007
3. Ling, V. B., Levi, E. E., Harrington, A. R., Zite, N. B., Rivas, S. D., Dalton, V. K., Smith, R., & Moniz, M. H. (2020). The cost of improving care: a multisite economic analysis of hospital resource use for implementing recommended postpartum contraception programmes. *BMJ quality & safety*, bmjqs-2020-011111. Advance online publication. https://doi.org/10.1136/bmjqs-2020-011111
4. Moniz, M. H., Chang, T., Davis, M. M., Forman, J., Landgraf, J., & Dalton, V. K. (2016). Medicaid Administrator Experiences with the Implementation of Immediate Postpartum Long-Acting Reversible Contraception. *Women's health issues : official publication of the Jacobs Institute of Women's Health*, *26*(3), 313–320. https://doi.org/10.1016/j.whi.2016.01.005
5. Moniz, M. H., McEvoy, A. K., Hofmeister, M., Plegue, M., & Chang, T. (2017). Family Physicians and Provision of Immediate Postpartum Contraception: A CERA Study. *Family medicine*, *49*(8), 600–606.
6. Moniz, M. H., Roosevelt, L., Crissman, H. P., Kobernik, E. K., Dalton, V. K., Heisler, M. H., & Low, L. K. (2017). Immediate Postpartum Contraception: A Survey Needs Assessment of a National Sample of Midwives. *Journal of midwifery & women's health*, *62*(5), 538–544. https://doi.org/10.1111/jmwh.12653
7. Moniz, M. H., Chang, T., Heisler, M., Admon, L., Gebremariam, A., Dalton, V. K., & Davis, M. M. (2017). Inpatient Postpartum Long-Acting Reversible Contraception and Sterilization in the United States, 2008-2013. *Obstetrics and gynecology*, *129*(6), 1078–1085. https://doi.org/10.1097/AOG.0000000000001970

Measure Number: 2903

Measure Title: Contraceptive Care – Most & Moderately Effective Methods

Measure Developer/Steward: HHS Office of Population Affairs

Reliability

- **Issue 1:** The following attachment was not provided: NQF_2903_Codes_2021-637453719019907247.xlsx
    - **Developer Response 1:** The codes were submitted along with the testing attachment.
- **Issue 2:** When does specification overlap with adequate demonstration of harmonizing with related measures?
    - **Developer Response 2:** OPA is submitting two other applications for NQF maintenance endorsement, which are complementary to NQF #2903. One of the applications is for NQF #2902 and focuses on use of most and moderately effective contraceptive methods in a key sub-population of women at risk of unintended pregnancy: postpartum women. The other application is for NQF #2904 and focuses on use of a sub-set of contraceptive methods, i.e., use of long-acting reversible contraception (LARC); the goal of this measure to monitor whether women have access to LARC methods as determined by whether any units report very low levels of LARC use (e.g., less than 1-2 percent). NQF #2903 uses the same code sets as NQF #2904 to calculate the denominator; the code

sets to calculate the NQF #2903 numerator also define the primary measure numerator for NQF #2902.
    - This measure is harmonized to the extent possible with the HEDIS Prenatal and Postpartum Care (PPC) measures, and some codes for live birth delivery, non-live birth, and pregnancy from the HEDIS PPC measures are utilized in the contraceptive care measures.
- **Issue 3:** It would have been helpful to have a more full description of the distribution of reliabilities.
    - **Developer Response 3:** We added histograms to show the distribution of reliability estimates at the group billing provider and facility levels at the end of this file in Figure 1. We did not provide histograms at the public health region or health plan levels because the number of units are small at these levels and reliability results for each unit at these levels were included in the application.

Validity

- **Issue 1:** Sensitivity results were less than desirable. For example, for contraceptive patch, it was 0.25. This is concerning that it is used to define numerator of this measure.
    - **Developer Response 1:** We agree that the low sensitivity for two elements (contraceptive patch and live birth in the last 2 months) are not ideal. However, for the contraceptive patch, the usage rate is very low and contributes to a very small proportion of total use of most or moderately effective methods. Four out of 423 patients (<1%) received a patch in our data according to the patient chart records. Thus, the impact of low sensitivity for this element should be minimal. In addition, because of the low rate of this method and relatively small total patient count in our data, the sensitivity estimate for this element may not reflect true sensitivity had we pulled a much larger sample of patient charts for data element validity.
    - For the "live birth in the last 2 months" element, we are aware of the potential challenge with capturing this particular element given that our only data source for chart reviews, title X clinics, do not have consistent access to live birth records within their systems. We would hypothesize significant improvements for this element within a fully-integrated hospital system where delivery information is captured consistently. OPA was not able to utilize hospital system data for the current application and will prioritize inclusion in the next re-endorsement application.
- **Issue 2:** Different levels of exclusion based on this variable were found among different data sources were concerning.
    - **Developer Response 2:** We noticed the differences on percentages of patients categorized into each category of the exclusion criteria, especially for "live birth in the last 2 months", across difference data sources and have investigated it. For PPFA, no patient was categorized into this category because PPFA does not provide delivery services or have access to delivery information, thus does not have patients' delivery records in their system. Comparing Washington (0.9%) and Iowa (4.6%) Medicaid data, one possible explanation is that Iowa may have more complete data regarding delivery/birth outcome, resulting in a higher percentage of "live birth in the last 2

months" and a lower percentage in "Unknown pregnancy outcome". Because the increased percentage in one category reduced the percentage in another, the total percentages of excluded women are comparable between these two states (5.9% for WA vs. 7.3% for IA), and thus the impact of these differences on measure calculation should be minimal.

- **Issue 3:** Concern on the exclusion of those who had a live birth in the last 2 months of the measurement year. This could potentially cause a lower incentive to achieve a successful score for these women. A simple date adjustment could be considered to avoid the exclusion of 2/12 months of data, as proposed for measure 2902. Additionally, no testing was conducted to assess how this exclusion criteria impacted the group level scores. It would be helpful to add such analysis to this submission.
    - **Developer Response 3:** OPA acknowledges the concern and will investigate the issue in future submissions. OPA does not have access to two consecutive years of data in order to provide this calculation in the current time frame.
    - While unable to calculate the impact for this submission, a small percentage of women (around 16%) gave birth in the last 2 months of the year and would thus have a reduced impact on the overall yearly rates; and we would not expect care providers to consider this measure-level exclusion or to perform services differently in a patient-centered setting. Further, the measure is not specified for use at the individual provider level (and is not for use as pay-for-performance) even should the service performance decrease for a small number of providers. In addition, the intention of the measure is to provide comparison of rates across measurement years to detect changes in clinic-or-higher-level performance over time. We do not expect this comparison to be affected as any impact of this exclusion on the rates should be consistent across years. Therefore, we think the impact of this exclusion should be minimal.
- **Issue 4:** Lack of risk adjustment
    - **Developer Response 4:** We acknowledge the concern on this issue but believe that risk adjustment isn't necessary for this measure. Although it is true that the rates differ by certain demographic characteristics, the underlying factors driving the differences are likely patient access to the service and/or patients' preference, which cannot be clearly or consistently attached to demographic characteristics. The differences in rates reflect differences in access of service by sociodemographic characteristics (financial and otherwise), and may not exist with equal access/quality of service in a patient-centered setting.
    - We calculate the measure by age group so that the measure scores for adolescents and adults for the purposes of quality improvement (QI), but not as a method of risk adjustment. Assessing the measure scores by age group is helpful because there are contraceptive programs and interventions focused specifically on adolescents and preventing teen pregnancy. Successful QI in these programs and interventions depend on the availability of contraceptive rates among adolescents.
- **Issue 5:** Unclear on patient-centeredness of this overall (face validity).

- **Developer Response 5:** Thank you for the feedback. OPA strongly believes that the goal of providing contraception should never be to recommend any one method or class of methods over women's individual choices. Thus, the contraceptive care measures are designed to encourage providers to offer those clients seeking contraception the full range of methods. Like NQF #2902 and #2904, #2903 is specified for use in administrative claims data, which has a limitation in that it does not contain information about patient preferences. Thus, OPA currently works to indirectly account for client choice and to strongly promote patient-centered contraceptive care through the Providing Quality Family Planning (QFP) guidelines [Gavin & CDC 2014]. Jointly published by OPA and CDC, QFP states that it is important that these contraceptive services are provided in a client-centered manner that treats each person as a unique individual with respect, empathy, and understanding, providing accurate, easy-to-understand information based on the client's self-identified needs, goals, preferences, and values. Clinics can utilize these guidelines with the contraceptive measures to deliver patient-centered contraceptive care.
- By design, NQF #2903 includes in its numerator all methods of contraception that require a prescription as well as sterilization. During measure calculation, these methods are treated as being of equal value. Hence, the numerator represents a wide range of methods from which clients can choose.
- Research indicates that patients receiving client-centered care may feel motivated to continue seeking reproductive health care for contraception and if they become pregnant, prenatal care and birth [Gomez 2017]. To broadly promote patient-centered reproductive health services, OPA delivers online education on this topic via a grant to the Reproductive Health National Training Center website (https://rhntc.org/resources/contraceptive-counseling-and-education-elearning). Recently updated in 2020, this on-demand training module is available to all providers.

- After NQF endorsed the contraceptive provision measures, OPA demonstrated its commitment to patient-centered contraceptive care by providing funding to the University of California San Francisco (UCSF) to develop a patient-reported outcome performance measure (PRO-PM) assessing the degree to which patient needs, values, and preferences are prioritized in the counseling encounter. After the initial year of funding, UCSF secured private funding to continue the project. Recently endorsed by NQF in December 2020 as the Person-Centered Contraceptive Counseling (PCCC) measure, this measure facilitates proper interpretation of the contraceptive provision measures by allowing organizations to observe variations in patient experience that occur with changes in provision of most or moderately effective contraception, including LARC methods. Health care providers can then ensure that increases in provision are not associated with a negative patient experience; ideally, improved provision would be linked to better patient experience. The PCCC's target population intersects with this measure's target population (e.g. ages 15-45 and assigned female at birth), but the PCCC is visit-specific. It is given to patients who have been identified as having received contraceptive counseling during their visit. A multi-organization partnership led by

UCSF and the National Association of Community Health Centers (NACHC) has started research to test the PCCC and developmental eCQM versions of NQF #2903 and #2904 in tandem use. Utilization of these two types of measures together can result in a more complete understanding of contraceptive care quality and help health care organizations to provide both access to a range of contraceptive methods and patient-centered counseling without coercion.

References

1. Gavin, L., Moskosky, S., Carter, M., Curtis, K., Glass, E., Godfrey, E., Marcell, A., Mautone-Smith, N., Pazol, K., Tepper, N., Zapata, L., & Centers for Disease Control and Prevention (CDC) (2014). Providing quality family planning services: Recommendations of CDC and the U.S. Office of Population Affairs. *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports*, *63*(RR-04), 1–54.
2. Gomez, A. M., & Wapman, M. (2017). Under (implicit) pressure: young Black and Latina women's perceptions of contraceptive care. Contraception, 96(4), 221–226. https://doi.org/10.1016/j.contraception.2017.07.007

Figure 1. Distribution of reliability among group billing providers who served ≥75 patients in Iowa Medicaid Enterprise, 2018

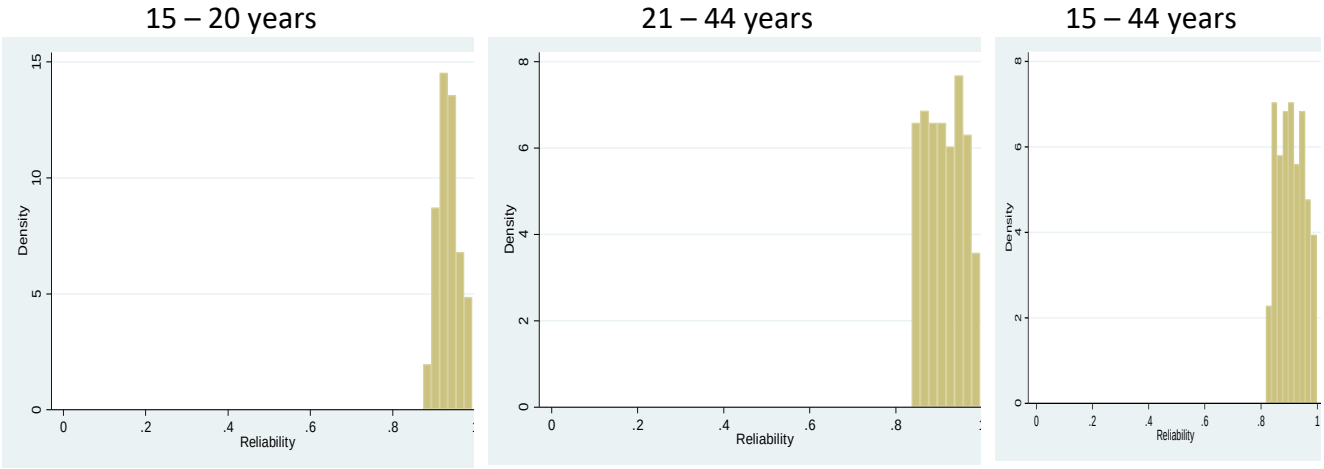


Figure 2. Distribution of reliability among 54 facilities who served ≥75 patients in PPFA, 2019

15 – 20 years 21 – 44 years 15 – 44 years

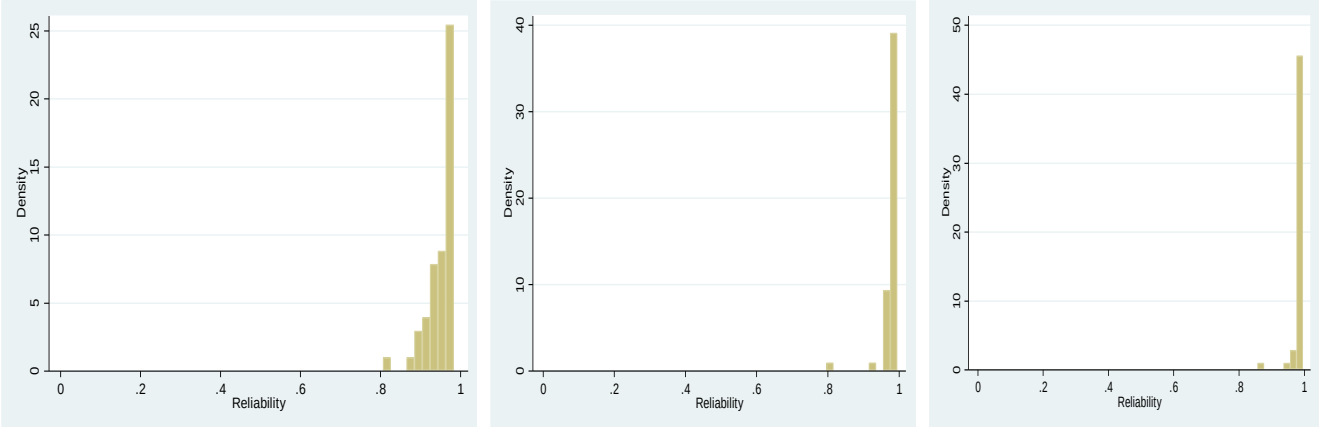
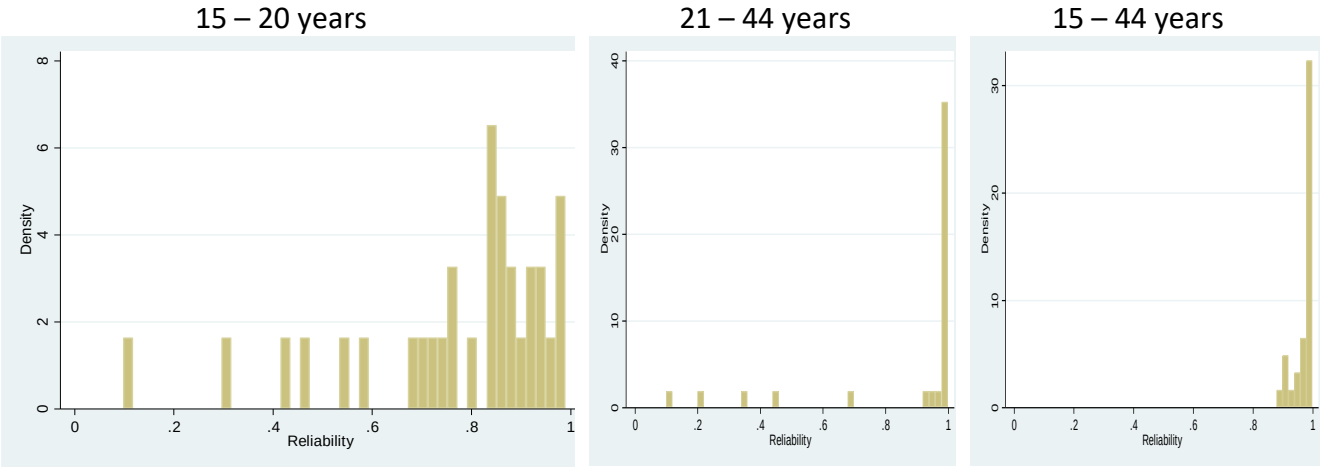


Figure 3. Distribution of reliability among 31 facilities who served ≥75 patients in NYP, 2018



Measure Number: 2904

Measure Title: Contraceptive Care – Access to LARC

Measure Developer/Steward: HHS Office of Population Affairs

Reliability

- **Issue 1:** Although the developer provided some testing results at group/practice level, the measure is not specified for use at group/practice level in the testing form. This is sensible but needs to be clear to measure users.
    - **Developer Response 1:** The measure is currently specified at facility, health plan, and public health region levels, as tested in the original application. In the current re-endorsement application, we tested the measure at these three levels as well as an additional group/practice level. If the measure receives re-endorsement at the group/practice level, then it will be specified at this level too, in addition to the three levels mentioned above. If that's the case, OPA as the measure steward, will educate

measure users on which levels the measure should be analyzed. The effort may include updating statements at the measure specification web page on the OPA website, where the published measure calculation programs are located; as well as presenting the updated measure specification at the annual Expert Work Group meetings, etc.

- **Issue 2:** Data dictionary not available: NQF_2904_Codes_2021.xlsx.
    - **Developer Response 2:** The codes were submitted along with the testing attachment.
- **Issue 3:** It was not clear to me how the calculated rates were used to compute the measure score. In the rationale, a 2% threshold was recommended. Was this the threshold used, i.e., less than 2% flagged a negative performance? This needs to be clarified.
    - **Developer Response 3:** This measure calculates the percentage of women aged 15-44 years at risk of unintended pregnancy that is provided a long-acting reversible method of contraception (i.e., implants, intrauterine devices or systems (IUD/IUS). We call this percentage the measure score or rate.
    - OPA designed this access measure to identify very low rates (less than 1-2%) of LARC use, which may signal barriers to LARC provision. Reporting units with scores less than 2% indicate that patients wishing to use LARC methods experience barriers to accessing this type of contraception. For this measure, we consider better quality to be a score within a defined interval (i.e., 2% or more).
- **Issue 4:** How reliably can one identify "at risk of unintended pregnancy" - a rhetorical question or an empirical one?
    - **Developer Response 4:** This measure relies on claims data and therefore, is not able to capture all women who are “at risk of unintended pregnancy” as we would like. We can only use information available in the claims system to identify these women, such as being infecund or being pregnant. We recognize this as a limitation of this measure. In collaboration with the University of California San Francisco, we are working on an eCQM version of the measure that uses EHR data to further capture women who are not at risk due to other reasons that are not available in the claims system. This developmental eCQM includes a new data element that enables patients to self-report their need for pregnancy prevention.
- **Issue 5:** “The measure steward, OPA recommends that the performance measure focus on low (rather than high) rates of use to evaluate women’s LARC access. For example, if a reporting entity has no or very few women using LARC (e.g., less than 2%), barriers restricting LARC access might be present and should be investigated.” No analysis was conducted on the reliability of being classified as a low outlier.
    - **Developer Response 5:** In Appendix F submitted with this application, we demonstrate a statistical tool that can be used to identify those units falling below a user-specified ‘floor’ value (e.g., 2% for this measure) with 95% confidence (while accounting for unit size and empirical distribution), to aid in assessments by quality improvement professionals.
- **Issue 6:** No description of the distribution of reliabilities was provided.
    - **Developer Response 6:** We added histograms to show the distribution of reliability estimates at the group billing provider and facility levels at the end of this file in Figure

1. We did not provide histograms at the public health region or health plan levels because the number of units are small at these levels and reliability results for each unit at these levels were included in the application.

Validity

- **Issue 1:** Sensitivity for two critical data elements is somewhat low. For example, for live birth data element, for age 21-44 group, the sensitivity is only 0.40.
    - **Developer Response 1:** We agree that the low sensitivity for "live birth in the last 2 months" is not ideal. We are aware of the potential challenge with capturing this particular element given that our only data source for chart reviews, title X clinics, do not have consistent access to live birth records within their systems. We would hypothesize significant improvements for this element within a fully-integrated hospital system where delivery information is captured consistently. OPA was not able to utilize hospital system data for the current application and will prioritize inclusion in the next re-endorsement application.
- **Issue 2:** Concern on the exclusion of those who had a live birth in the last 2 months of the measurement year. This could potentially cause a lower incentive to achieve a successful score for these women. A simple date adjustment could be considered to avoid the exclusion of 2/12 months of data, as proposed for measure 2902. Additionally, no testing was conducted to assess how this exclusion criteria impacted the group level scores. It would be helpful to add such analysis to this submission.
    - **Developer Response 2:** OPA acknowledges the concern and will investigate the issue in future submissions. OPA does not have access to two consecutive years of data in order to provide this calculation in the current time frame.
    - While unable to calculate the impact for this submission, a small percentage of women (around 16%) gave birth in the last 2 months of the year and would thus have a reduced impact on the overall yearly rates; and we would not expect care providers to consider this measure-level exclusion or to perform services differently in a patient-centered setting. Further, the measure is not specified for use at the individual provider level (and is not for use as pay-for-performance) even should the service performance decrease for a small number of providers. In addition, the intention of the measure is to provide comparison of rates across measurement years to detect changes in clinic-or-higher-level performance over time. We do not expect this comparison to be affected as any impact of this exclusion on the rates should be consistent across years. Therefore, we think the impact of this exclusion should be minimal.
- **Issue 3:** Different levels of exclusion based on this variable were found among different data sources.
    - **Developer Response 3:** We noticed the differences on percentages of patients categorized into each category of the exclusion criteria, especially for "live birth in the last 2 months", across difference data sources and have investigated it. For PPFA, no patient was categorized into this category because PPFA does not provide delivery services or have access to delivery information, thus does not have patients' delivery records in their system. Comparing Washington (0.9%) and Iowa (4.6%) Medicaid data,

one possible explanation is that Iowa may have more complete data regarding delivery/birth outcome, resulting in a higher percentage of "live birth in the last 2 months" and a lower percentage in "Unknown pregnancy outcome". Because the increased percentage in one category reduced the percentage in another, the total percentages of excluded women are comparable between these two states (5.9% for WA vs. 7.3% for IA), and thus the impact of these differences on measure calculation should be minimal.

- **Issue 4:** Lack of risk adjustment.
    - **Developer Response 4:** Because NQF #2904 focuses on identifying very low rates of LARC provision which may indicate barriers to access, we believe that risk adjustment is not justified. By design, this measure is a floor measure to help entities determine if any LARC provision occurs at all. Reporting units with NQF #2904 scores less than 2% may have barriers restricting LARC access. These entities should be further evaluated to determine which barriers are present and to develop a plan to address them. Units with scores greater than or equal to 2% should be considered as providing clients desiring LARC with access to these methods. OPA also strongly recommends that measure users also investigate very high NQF #2904 scores that are outliers (i.e., rates that approach 100%). It is extremely important to ensure the provision of patient-centered contraceptive counseling and that women are not being coerced into receiving LARC methods.
- **Issue 5:** Given the emphasis on using this measure to identify very low rate to uncover potential barriers for access to LARC, it is not clear how to interpret the rate differences among entities when rates were not lower than 2%.
    - **Developer Response 5:** When entities within a health system have rates for #2904 which are higher than 2%, one can assume that patients have access to LARC, and no additional interpretation is needed. Although NQF #2904 is designed to identify very low rates, OPA also strongly recommends that measure users also investigate very high #2904 scores that are outliers (i.e., rates that approach 100%). It is extremely important to ensure the provision of patient-centered contraceptive counseling and that women are not being coerced into receiving LARC methods. A range of contraceptive preferences is expected. It is vital that women who wish to use contraception have the full range of methods available to them, and not only one type of method.

- **Issue 6:** The developer should provide clear guidance to measure users on how to interpret the results, particularly when they may intend to compare rates across settings. For example, mean rate for facility in PPFA was 0.135 while mean rate for facility in NYP was 0.072. For this measure, typical better or worse than average performance may not be an appropriate reporting method.
    - **Developer Response 6:** When comparing rates across entities, OPA recommends that the performance measure focus on whether scores are below 2%. Two separate entities reporting measure rates greater than 2% indicate that access to LARC exists for both client populations and comparison of the mean is not needed.

- **Issue 7:** No analysis of missing data.
    - **Developer Response 7:** We did not conduct analysis of missing data because the data source for this measure is claims data. Due to the nature of claims data (i.e., for billing purposes), there is typically very little missing data. Further, it is difficult to ascertain when claims data is or is not missing and thus we were not able to conduct such analysis.
- **Issue 8:** Unclear on patient-centeredness of this overall (face validity).
    - **Developer Response 8:** OPA strongly believes that the goal of providing contraception should never be to recommend any one method or class of methods over women's individual choices. Thus, the contraceptive care measures are designed to encourage providers to offer those clients seeking contraception the full range of methods. Like #2902 and #2903, #2904 is specified for use in administrative claims data, which has a limitation in that it does not contain information about patient preferences. Thus, OPA currently works to indirectly account for client choice and to strongly promote patient-centered contraceptive care through the Providing Quality Family Planning (QFP) guidelines [Gavin & CDC 2014]. Jointly published by OPA and CDC, QFP states that it is important that these contraceptive services are provided in a client-centered manner that treats each person as a unique individual with respect, empathy, and understanding, providing accurate, easy-to-understand information based on the client's self-identified needs, goals, preferences, and values. Clinics can utilize these guidelines with the contraceptive measures to deliver patient-centered contraceptive care.
    - Research indicates that patients receiving client-centered care may feel motivated to continue seeking reproductive health care for contraception and if they become pregnant, prenatal care and birth [Gomez 2017]. To broadly promote patient-centered reproductive health services, OPA delivers online education on this topic via a grant to the Reproductive Health National Training Center website (https://rhntc.org/resources/contraceptive-counseling-and-education-elearning). Recently updated in 2020, this on-demand training module is available to all providers.
    - After NQF endorsed the contraceptive provision measures, OPA demonstrated its commitment to patient-centered contraceptive care by providing funding to the University of California San Francisco (UCSF) to develop a patient-reported outcome performance measure (PRO-PM) assessing the degree to which patient needs, values, and preferences are prioritized in the counseling encounter. After the initial year of funding, UCSF secured private funding to continue the project. Recently endorsed by NQF in December 2020 as the Person-Centered Contraceptive Counseling (PCCC) measure, this measure facilitates proper interpretation of the contraceptive provision measures by allowing organizations to observe variations in patient experience that occur with changes in provision of most or moderately effective contraception, including LARC methods. Health care providers can then ensure that increases in provision are not associated with a negative patient experience; ideally, improved provision would be linked to better patient experience. The PCCC's target population intersects with this

measure's target population (e.g., ages 15-45 and assigned female at birth), but the PCCC is visit-specific. It is given to patients who have been identified as having received contraceptive counseling during their visit. A multi-organization partnership led by UCSF and the National Association of Community Health Centers (NACHC) has started research to test the PCCC and a developmental eCQM version of NQF #2904 in tandem use. Utilization of these two types of measures together can result in a more complete understanding of contraceptive care quality and help health care organizations to provide both access to a range of contraceptive methods and patient-centered counseling without coercion.

References

1. Gavin, L., Moskosky, S., Carter, M., Curtis, K., Glass, E., Godfrey, E., Marcell, A., Mautone-Smith, N., Pazol, K., Tepper, N., Zapata, L., & Centers for Disease Control and Prevention (CDC) (2014). Providing quality family planning services: Recommendations of CDC and the U.S. Office of Population Affairs. *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports*, *63*(RR-04), 1–54.
2. Gomez, A. M., & Wapman, M. (2017). Under (implicit) pressure: young Black and Latina women's perceptions of contraceptive care. Contraception, 96(4), 221–226. https://doi.org/10.1016/j.contraception.2017.07.007

Figure 1. Distribution of reliability among group billing providers who served ≥75 patients in Iowa Medicaid Enterprise, 2018

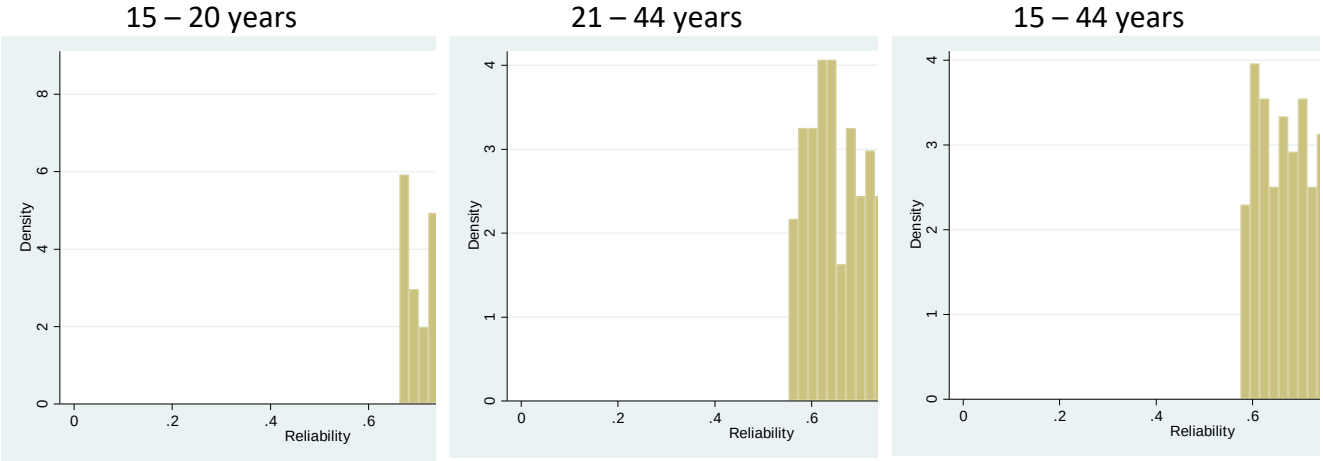


Figure 2. Distribution of reliability among 54 facilities who served ≥75 patients in PPFA, 2019

15 – 20 years 21 – 44 years 15 – 44 years

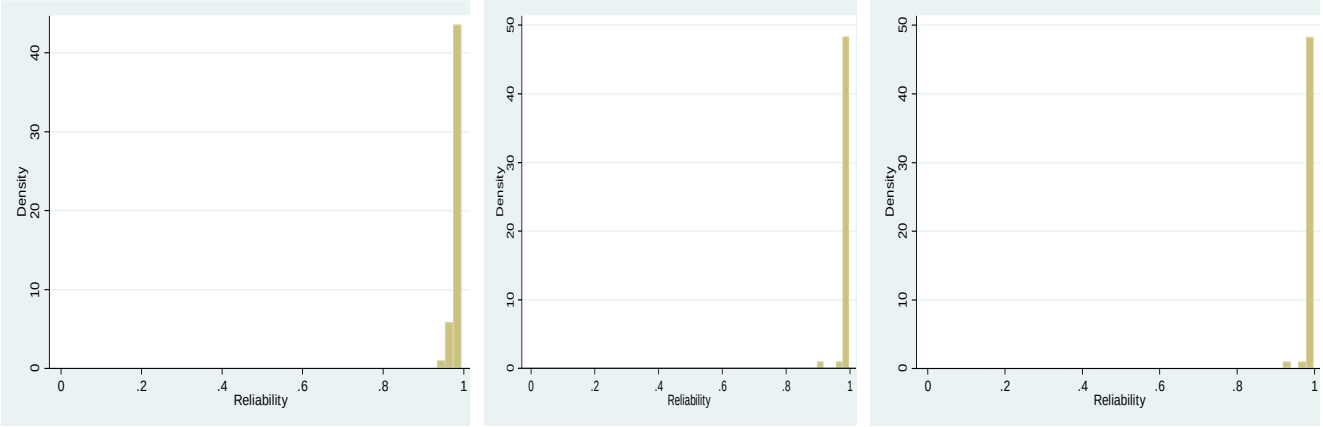
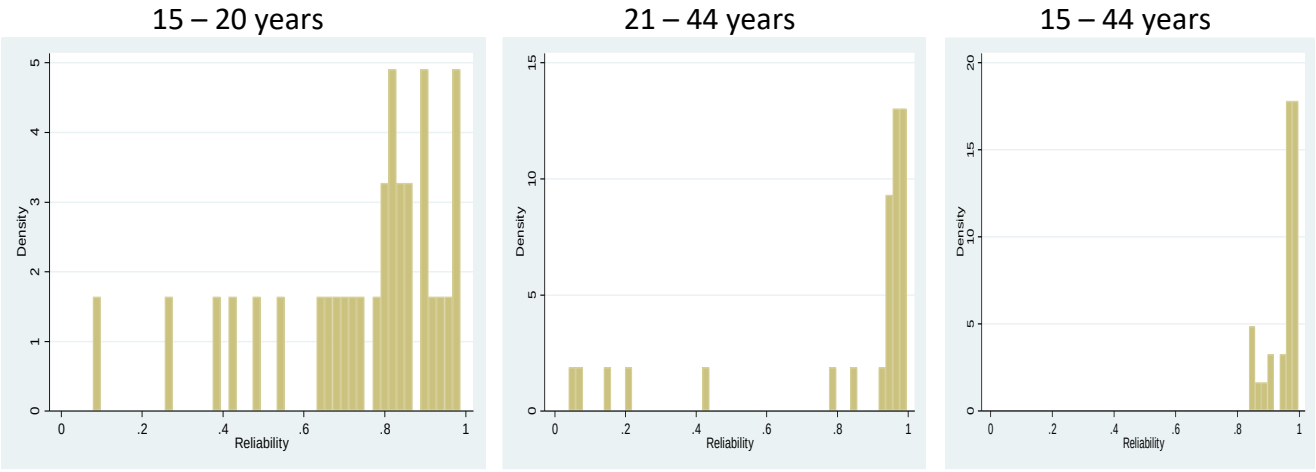


Figure 3. Distribution of reliability among 31 facilities who served ≥75 patients in NewYork-Presbyterian Hospital system, 2018



Measure Number: 3501e

Measure Title: Hospital Harm - Opioid-Related Adverse Events

Measure Developer/Steward: IMPAQ International LLC

Reliability

- **Issue 1:** clarify how the measure can differentiate between use of Naloxone as an indicator of opioid-related adverse events vs. other uses following or in combination with opioid use
    - **Developer Response 1:** We appreciate the panel member's comment and agree that it is important that the eCQM, as currently specified, not detect false positives. Absent a prior opioid administration, naloxone may be given inside of the operating room as part of an anesthesia plan or may be used for purposes other than opioid reversal, such as off-label use for opioid-induced pruritus. We incorporated feedback from the NQF Patient Safety Standing Committee (2019 Spring Cycle) into measure refinement, and made several important changes. First, we have added a qualifying restriction that

requires a hospital-administered opioid to precede the naloxone administration in order to count as a numerator event. Second, the measure requires that naloxone must be administered within 12 hours after the opioid administration in order to count the naloxone administration as a numerator event. Third, when qualifying a potential numerator event, the measure excludes naloxone administration occurring in the operating room setting.

- We conducted empirical tests to examine whether numerator cases identified by the measure are true positives. In the chart review (or parallel-form comparison) process we instructed clinical abstractors to extract both indications for and patient subsequent responses to the naloxone administration. We found that the predominant rationale for subsequent naloxone administration was that patients were somnolent or unresponsive, with the second mostly cited reason being opiate reversal. In terms of patient responses to naloxone administration, we found that the most frequently documented was: patient showed clear signs of response to naloxone administration. This qualitative evidence solidifies the evaluation of measure logic and suggests that the measure can correctly predict a true positive.

- **Issue 2:** Is there a way to directly identify an opioid-related adverse event from the EHR without using Naloxone as the signal of this measure
  - **Developer Response 2:** Naloxone administration has been used in studies of computerized adverse drug event surveillance as an indicator of severe opioid-related adverse events.[1,2] By encouraging hospitals to implement evidence-based practices such as routine patient monitoring for potential adverse effects of opioids, this eCQM can lead to better quality of care associated with excessive opioid administration in the hospital setting.[3,4]
- **Issue 3:** Why the measure only excludes naloxone administration occurring in the operating room setting when qualifying the potential harm event, and why not also excludes naloxone administration in PACU or ICU?
  - **Developer Response 3:** We appreciate the panel member's comment and we acknowledge that naloxone administration in these other settings may be part of the anesthesia plan. However, during the stage of measure feasibility testing, we found that temporary location documentation in the EHR is limited and inconsistent, particularly among the less advanced EHRs. Even for the more advanced EHRs, such as Epic, documentation of temporary locations inside of hospital in the structured fields is not consistent across hospitals. Facing this practical constraint that applies globally to a great majority of EHRs, we have thus decided to only include the operating room in the measure specification. As EHR technology advances, we will revisit the measure and re-evaluate the feasibility of excluding naloxone administrations in other locations.
- **Issue 4:** Does the measure value set include oral naloxone or IV only?
  - **Developer Response 4:** We appreciate the panel member's question and clarify that the measure value set does not include oral naloxone. Our clinical rationale for this decision is twofold. First, the intent of the measure is to capture events when an opioid antagonist is needed to reverse an opioid overdose that occurs in the hospital setting. Hence, it would not be appropriate to treat the patient with an oral form (which is slow

acting), but rather the fast-acting nasal spray or injectable solutions. Second, oral naloxone forms exist, yet they are used (in low doses) on occasion to reverse opioid-induced constipation rather than overdose. Thus, these would not be appropriate to be considered as an opioid antagonist for the measure. We recognize that there are oral and sublingual forms of medications that include naloxone as one of the ingredients, such as Suboxone, which is a combination of naloxone and buprenorphine. However, this combination of medications is to treat opioid addiction and would not be appropriate as opioid antagonists for the measure.

- **Issue 5:** Does the measure use nursing notes for calculation? And how nursing notes are "validated?"
    - **Developer Response 5:** We appreciate the opportunity to clarify two points. First, the measure, as currently specified, does not use nursing notes for calculation, in that nursing notes are mostly unstructured. During the chart review process, we instructed clinical abstractors to look at nursing notes and identify the indication for and subsequent patient response to the naloxone administration. These notes help us understand whether or not the measure identifies false positives and if it can correctly predict a true positive. Second, we did not validate/audit nursing notes during chart reviews as that is outside the scope of analysis. Nursing notes are, however, subject to the test site's auditing procedure, though they may not be scrutinized with forensic exactitude.
- **Issue 6:** Given the extremely low event rates <1%, its unclear if the reliability testing included any events in the random sample of cases.
    - **Developer Response 6:** We appreciate the opportunity to clarify two points. First, to assess the data element reliability, we used two methods. Method 1 calculated the rate of missing or erroneous values for every critical data element needed for measure implementation, and method 2 computed Cohen's Kappa for each of the six implementation test sites to gauge inter-rater agreement on the critical data elements (rater A is the EHR and rater B is the clinical abstractor). Method 1 was based on the full sample, while method 2 was based on the six randomly selected sub-samples, each of which consisted of 100 encounters, and was drawn from the full sample from a given implementation test site. In five of the six sub-samples, we included all numerator encounters identified in the full sample, and sixth sub-sample contained 50 numerator encounters.
- **Issue 7:** Need clarity on the definition of "erroneous" and "gold standard". Further, explain how the following is plausible, "Error rates are 0% and Kappas are 0.98."
    - **Developer Response 7:** We appreciate the panel member's comment. First, we define "erroneous" as EHR data showing values inconsistent with the variable format or data values that are wrong for obvious reasons. For example, one of the critical data elements is the timestamp of opioid administration. For such a variable, we would expect its value to be stored and displayed in any conventional timestamp format, such as 3/3/2021 12:26:00 PM. Another example is patient age equal to 999. Second, we define "gold standard" as data extracted from chart reviews, which is a common tool used in many settings, such as Medicare Advantage Organizations using chart reviews to

improve the accuracy of their risk-adjusted payments. We, however, recognize that its definition is subject to confirmation bias. Third, the error rate being zero indicates that the critical data elements are reliably captured and stored in the EHR, while $\kappa = 0.98$ indicates that what EHR captured and stored are nearly always consistent with what one would see in the patient's medical chart. Hence, the former is focused on the EHR alone, while the latter focuses on the comparison between two data sources. The two statistics are not competing against each other.

Validity

- **Issue 1:** The measure score level validity testing is no different than the data element level validity testing
    - **Developer Response 1:** We appreciate panel members' comments on the suitability of measure score level validity testing method, and we would like to clarify two points. First, we agree with panel members that an orthodox approach to assessing the score level validity is the measure construct validity, or the extent to which the measure generates estimates that are consistent with a construct or conceptual framework regarding how safe care is produced and defined. To assess the construct validity, measure developers typically use the Spearman rank correlation, quantifying the relationship between the measure of interest and those of a single underlying concept. This approach is commonly used in claims-based measures. We, however, caution that this method will be less informative given the number of test sites (a total six) that participated in measure testing. Small sample size will yield a spurious p-value. Second, we did not compare the measure performance rate at the accountable entity level to other external measures of quality because, for measures that count harm events without other statistical manipulations, the confirmation that measure logic is accurately capturing true harm events is an appropriate method for assessing the validity of measure score. Third, not a clarification but an agreement that we will re-evaluate the measure score level validity using the conventional approaches after measure implementation has occurred and there is data available from more hospitals.
- **Issue 2:** Across the six test sites the measure performance rate ranged from 0.11% to 0.45%, but whether any of the sites were statistically different from the average or each other was not shown
    - **Developer Response 2:** We appreciate the panel member's comment and the opportunity to make one clarification. We understand the importance of showing if the differences in measure performance rate across accountable entities are statistically different, but are at the same time cautious about the credence of p-value estimated from a small sample. Of note, there were a total of six hospitals that participated in measure implementation testing.
- **Issue 3:** The necessity of risk adjustment in this measure
    - **Developer Response 3:** We appreciate panel members' comments on the utility of risk adjustment in the measure, and recognize that in many cases accounting for the patient mix and service mix is essential in outcome measures. We thus thank the opportunity to make the following clarifications. First and foremost, we agree with panel members'

standpoint on the utility of risk adjustment at a high level. Second, we did not apply risk adjustment in the measure, in that 1) there is strong evidence showing that most instances of severe over-sedation requiring naloxone for reversal can be avoided by following best practices; 2) the most common cause of ORAE is hospital administration of excessive doses or over-sedation and inadequate monitoring even when certain patients may require higher doses to achieve pain control or are more sensitive to opioids (depending on their age, sex, and weight); and 3) the dosing of opioids and the intensity of patient monitoring is under the control of providers in hospitals, and risk can be minimized by following best practices. Third, the inappropriateness for not subjecting some accountable entities to the same level of scrutiny as other accountable entities simply because the patient mix differs. ORAEs should be avoidable regardless of patient risk, especially when the opioid was given after patients have arrived at the hospital.

Other General Comments

- We are grateful for panel members' comments on the measure in various aspects, and the suggestions to better demonstrate the measure's scientific rigor. We agree with the suggestion that, when data from more hospitals (with different size, geographic location, urban/rural representation, academic and non-academic status) are collected, we will re-evaluate the measure score level reliability and validity using approaches that are frequently employed during widespread implementation.

References

1. Nwulu, U., Nirantharakumar, K., Odesanya, R., McDowell, S. E., & Coleman, J. J. Improvement in the detection of adverse drug events by the use of electronic health and prescription records: an evaluation of two trigger tools. Eur J Clin Pharmacol. 2013;69(2), 255-259.
2. Eckstrand, J. A., Habib, A. S., Williamson, A., Horvath, M. M., Gattis, K. G., Cozart, H., & Ferranti, J. Computerized surveillance of opioid-related adverse drug events in perioperative care: a cross-sectional study. Patient Saf Surg. 2009;3(1), 18.
3. Practice Guidelines for the Prevention, Detection, and Management of Respiratory Depression Associated with Neuraxial Opioid Administration. Anesthesiology. 2009;110(2):218-230.
4. Lee LA, Caplan RA, Stephens LS, et al. Postoperative opioid-induced respiratory depression: a closed claims analysis. Anesthesiology. 2015;122(3):659-665.

Measure Number: 3621

Measure Title: Composite weighted average for 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan, CT Head/Brain without contrast/single phase scan)

Measure Developer/Steward: American College of Radiology (ACR)

Reliability

- **Issue 1:** The developer indicated that the measure is specified for a facility but did not provide any facility-level reliability testing.

    - **Developer Response 1:** Facility-level data for this measure is provided below using the relevant sections from the testing form.

DATA/SAMPLE USED FOR FACILITY LEVEL TESTING

What type of data was used for testing?

Measure Specified to Use Data From:	Measure Tested with Data From:
☐ abstracted from paper record	☐ abstracted from paper record
☐ claims	☐ claims
☒ registry	☒ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.1. If an existing dataset was used, identify the specific dataset.

The American College of Radiology (ACR) used data from their National Radiology Data Registry (NRDR) Dose Index Registry (DIR). The primary participants ("target population") are hospital radiology departments (inpatient/outpatient), radiology groups and free-standing imaging centers.

1.3. **What are the dates of the data used in testing**? January 1, 2017 – December 1, 2020.

1.4. What levels of analysis were tested? Hospital/facility/agency (Group analysis previously provided)

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)?

The testing sample comprised all facilities that submitted data to ACR NRDR DIR for this measure. The sample consisted of 2,863 hospitals/imaging facilities. The eligible population for this measure (i.e. the denominator) includes all submitted eligible records (CT Abdomen-pelvis with IV contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan). There are no exclusions with this measure.

Table 1. Number of facilities that submitted data for this measure.

	Composite Weighted Average of all 3	CT Abdomen-pelvis with IV contrast/single phase scan	CT Chest without contrast/single phase scan	CT Head/Brain without contrast/single phase scan
All Years	2,893	2,721	2,782	2,743
2017	2,148	1,916	1,937	1,929
2018	2,390	2,141	2,182	2,132
2019	2,428	2,105	2,220	2,112
2020	2,386	2,090	2,204	2,079

1.6 How many and which patients were included in the testing and analysis (by level of analysis and data source)?

A total of 50,356,186 patients were eligible to be included in this testing. Patients included both male and female of all ages with various indications for the exams in each measure exam category. The registry categorizes data by study description and covers any indication that may be associated with the procedure.

Table 2. Eligible patients and reported patients

	Composite Weighted Average of all 3		CT Abdomen-pelvis with IV contrast/single phase scan		CT Chest without contrast/single phase scan		CT Head/Brain without contrast/single phase scan	
	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported
All Years	50,356,186	50,096,936	17,229,385	17,132,224	7,397,579	7,364,059	25,723,543	25,594,982
2017	10,546,008	10,525,572	3,514,958	3,509,303	1,383,514	1,380,782	5,646,213	5,634,170
2018	12,845,656	12,785,646	4,353,122	4,331,936	1,781,313	1,774,451	6,709,691	6,677,730
2019	14,174,013	14,087,765	4,873,574	4,838,446	2,157,910	2,146,430	7,140,965	7,101,326
2020	12,790,509	12,697,953	4,487,731	4,452,539	2,074,842	2,062,396	6,226,674	6,181,756

1.7 If there are differences in the data or sample used for different aspects of testing (e.g.,

reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

There are no differences in the data or sample used for different aspects of testing.

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

No patient-reported or social risk factors data are available for this measure.

2a2. RELIABILITY TESTING

2a2.1. What level of reliability testing was conducted? Performance measure score using signal-to-noise ratio analysis.

2a2.3. What were the statistical results from reliability testing?

Using the parameter estimates from the beta-binomial model, we computed and aggregated reliability scores for each year. Please see **Table 3** for the results.

Table 3. Reliability score statistics by year by component.

Year	*Number of Facilities*	*25th percentile*	*Reliability median*	*75th percentile*	*Reliability mean*	*Lower Confidence Limit (minimum)*	*Upper Confidence Limit (maximum)*
Composite Weighted Average of All 3							
2017	2150	.99997	.99999	1.00000	.99994	.99994	.99995
2018	2390	.99998	.99999	1.00000	.99996	.99996	.99997
2019	2430	.99997	.99999	1.00000	.99996	.99995	.99996
2020	2386	.99998	.99999	1.00000	.99995	.99995	.99996
ALL	2,893	.99998	.99999	1.00000	.99995	.99995	.99996
Component 1: CT Abdomen-pelvis							
2017	1920	.99984	.99996	.99999	.99974	.99971	.99977
2018	2146	.99985	.99996	.99999	.99977	.99974	.99979

Year	Number of Facilities	25th percentile	Reliability median	75th percentile	Reliability mean	Lower Confidence Limit (minimum)	Upper Confidence Limit (maximum)
2019	2116	.99987	.99997	.99999	.99979	.99976	.99981
2020	2099	.99989	.99997	.99999	.99983	.99980	.99985
ALL	2,090	.99986	.99996	.99999	.99978	.99977	.99979
Component 2: CT Chest							
2017	1939	.99986	.99996	.99999	.99983	.99981	.99984
2018	2184	.99990	.99997	.99999	.99987	.99986	.99988
2019	2224	.99992	.99997	.99999	.99989	.99988	.99990
2020	2205	.99991	.99997	.99999	.99988	.99986	.99989
ALL	2,204	.99990	.99997	.99999	.99987	.99986	.99987
Component 3: CT Head/Brain							
2017	1953	.99936	.99987	.99996	.99869	.99852	.99886
2018	2162	.99941	.99988	.99996	.99888	.99873	.99902
2019	2147	.99913	.99984	.99995	.99845	.99825	.99864
2020	2122	.99894	.99977	.99993	.99813	.99790	.99837
ALL	2,079	.99922	.99984	.99995	.99853	.99844	.99863

- **2a2.4 What is your interpretation of the results in terms of demonstrating reliability**?

  A reliability score of 0.7 is considered a reasonable minimum threshold for reliability. Based on the mean reliability scores of 0.9995 for the composite measure, this measure is considered reliable. Each individual measure component is producing consistent and accurate results, as is the composite measure using the current weighting algorithm. The measures as defined reliably identify variability in performance across facilities.

- **Issue 2:** Near perfect reliability as measured by the SNR. Especially with the very large sample size, a split sample reliability analysis and a 'stability of classification' analysis would have been illuminating.
    - **Developer Response 2:** It is true that our sample sizes are very large, and this results in reliability of close to 1 at all levels of performance for an average sample, at both the facility and group level. While we did not perform the analyses raised here, we did assess reliability for sample sizes as small as 20: reliability is .7 or higher across performance levels.

Validity

- **Issue 1:** It appears the developer did not convene a panel to establish face validity but provided access to various consensus reports.
    - **Developer Response 1:** The ACR convened a formal panel that recently completed a face validity survey for this measure. The expert panel members consisted of medical physicists, radiologists, a value-based purchasing surveyor, and a patient:
        - Missy Danforth (Value-based Purchasing Surveyor) Washington, DC
        - Demetrios Giannikopoulos (Patient) Ellicott City, MD
        - Chad Dillon (Medical Physicist) Hoover, AL
        - Kyle Jones (Medical Physicist) Houston, TX
        - Alexander Towbin, MD (Physician) Cincinnati, OH
        - David Jordan (Medical Physicist) Cleveland, OH
        - Doug Kitchin, MD (Physicist) Middleton, WI
        - Olga Brook, MD (Physician) Boston, MA
        - Kimberly Applegate, MD (Physician) Zionsville, IN
        - Randell Kruger, PhD (Medical Physicist) Marshfield, WI
        - Beth Schueler, MD (Physician) Rochester, MN
        - Loretta Johnson (Medical Physicist) Birmingham, AL
        - Nadja Kadom, MD (Physician) Atlanta, GA
        - Donald Frush, MD (Physician) Durham, NC
        - William Breeden (Medical Physicist) Indianapolis, IN
        - James Tomlinson (Medical Physicist) Ann Arbor, MI
        - Tyler Fisher (Medical Physicist) Signal Hill, CA
        - Clinton Jokerst, MD (Physician) Scottsdale, AZ
        - Tony Seibert, MD (Physician) Sacramento, CA
        - Eric Rubin, MD (Physician) Upland, CA
        - David Seidenwurm, MD (Physician) Sacramento, CA
    - The panel was asked three questions:
        - Do you think that monitoring radiation dose indices from clinical CT exams is a good and worthwhile activity for advancing or maintaining safety and quality?
        - Is this measure as described a reasonable and appropriate way to assess performance quality of a facility or practice with regards to dose optimization?
        - Will the scores obtained from the measure as specified reasonably differentiate clinical performance across providers, and separate the high performers from

the low performers?

- 95% of the panel (20 members) agreed that monitoring radiation dose indices from clinical CT exams is a good and worthwhile activity for advancing or maintaining safety and quality. One panel member mentioned that radiation dose indices are particularly important when associated with adequate image quality. Another commented that radiation dose indices are a well-recognized method to compare site CT patient dose indices to a national benchmark. Two members mentioned that patient size must be considered as well. The ACR strongly agrees with the panel comments. This measure does take patient size into consideration and is one of the data elements collected automatically during data transmission to the registry. The one panel member that did not agree with this statement did not leave a comment for their response.
- 71% of the panel (15 members) agreed that the measure as described is a reasonable and appropriate way to assess performance quality of a facility or practice with regards to dose optimization. A member stated that they agreed but that it is more important to ensure the image quality is sufficient to make an accurate diagnosis; the ACR strongly agrees that the image quality is vital for an accurate diagnosis. Unfortunately, there are no standards for quantifying image quality at this time. Measuring size-specific exam level DLPs accommodates this concern to some extent as it allows for dose index differentials to obtain diagnostic quality images across patients of different sizes. 29% (6 members) while not specifically stating that the measure was not reasonable or appropriate, did not agree that measure is the *best* way to assess performance quality; we believe this is reflective of the general desire for advances in quantitative image quality assessment, which is beyond the scope of this measure. Two members commented that the scan lengths can vary and that CTDI or size-specific dose estimate (SSDE) would be the better metric. We do not disagree that SSDE is a better metric of size-specific dose; however, most scanners do not report SSDE and all scanners report DLP. Our use of diagnostic reference levels based on size-specific DLPs is a compromise allowing all facilities to be able to use this measure and improve their performance. Another member stated that the measure should also include percent of DLP at or below the size-specific mean reference level, but this suggestion is not aligned with the standardized methodology for calculating diagnostic reference levels, provided in Publication 135 of the International Commission on Radiological Protection, entitled "Diagnostic reference levels in medical imaging." but this suggestion is not aligned with the standardized methodology for calculating diagnostic reference levels, provided in Publication 135 of the International Commission on Radiological Protection, entitled "Diagnostic reference levels in medical imaging." We provide comparisons at multiple levels to facilitate process improvement. For an accountability measure, in the interest of parsimony, we use the 75th percentile (diagnostic reference level) benchmarks rather than the 50th percentile (achievable dose) benchmark or the mean, as DRLs are more widely used across programs in radiology departments. Also using DRLs for accountability is supported across US and international organizations:
  - NCRP Report No. 172, *Reference Levels and Achievable Doses in Medical and*

Dental Imaging: Recommendations for the United States (National Council on Radiation Protection and Measurements)
        - ACR–AAPM–SPR PRACTICE PARAMETER FOR DIAGNOSTIC REFERENCE LEVELS AND ACHIEVABLE DOSES IN MEDICAL X-RAY IMAGING (ACR, AAPM, SPR)
        - ICRP Publication 135, *Diagnostic reference levels in medical imaging* (International Commission on Radiological Protection)
    - 62% of the panel (13 members) agreed that the scores obtained from the measure would differentiate clinical performance across providers. One panelist stated that using DRLs was a positive step forward in this effort. 38% of the panel (8 members) did not agree. Three panelists mentioned the age of the CT scanner as an important variable in quality; older machines will need higher doses to obtain the same image quality, therefore low performers may be more related to how the old the equipment is. The ACR thinks this is an important point to stress. Patients want the best quality possible on their exams, and if providers are subjecting patients to higher doses because of old equipment, it is vital to capture this information. Another panelist commented that DRLs were not meant to differentiate performance. The ACR agrees that using DRLs alone is not an appropriate way to calculate performance, as it's not a measure or estimate of actual patient radiation dose, but it is closely related to the doses received by patients. DLP is a measure of radiation output received and experienced by patients and not simply documentation of whether DLP was recorded. The ACR also collects patient size information so dose estimates can be adjusted accordingly. Providing comparative data across exam types to a physician or site will help adjust imaging protocols to obtain diagnostic images using the lowest reasonable radiation dose. This measure collects the CT scanner radiation output specific to a patient and exam and compares the actual dose indices to benchmarks for similarly sized patients and similar exam types.
- **Issue 2:** I'm not sure of the methods the developer used to collect, review and evaluate the literature they did review.
    - **Developer Response 2:** The literature provided was curated by ACR Senior Advisor for Medical Physics, Dustin Gress. Mr. Gress is a diagnostic and nuclear medical physicist, board certified by the American Board of Radiology and the American Board of Science in Nuclear Medicine. He has approximately 14 years of clinical experience, spending roughly half in private practice—supporting upwards of 200 client facilities ranging from academic hospitals to rural standalone imaging clinics—and the other half in a high-volume academic cancer hospital. Mr. Gress selected the submitted literature based on his experience as a medical physicist, in order to demonstrate broad, both national and international, expert consensus support for medical imaging practices to monitor their use of radiation dose in patient imaging and compare their performance to available benchmarks. The organizations whose documents are referenced are the standard bearers in their space and are widely followed. ICRP guidance is followed by EU nations and others around the world for national policymaking and clinical practice guidance; the NCRP is similarly regarded in the US. ACR and AAPM are professional organizations

that define standards of care for medical physics and the radiological professions in the US.

- **Issue 3:** It appears measure stewards have access to perform measure score validity testing. I would recommend that they do this analysis and resubmit.
    - **Developer Response 3:** The ACR appreciates this suggestion and will perform empirical validity for measure maintenance per NQF testing requirements.
- **Issue 4:** There is no double check on whether the ACR NRDR DIR participants check on the quality of the data e.g., what does the developer do to ensure data accuracy? The developer indicates that "no missing data was found through testing, nor would missing data be expected to occur in the future." No testing results presented.
    - **Developer Response 4:** No missing data is technically possible given the direct submission from the software in the scanners to the registry. The standard required content for data transmission to the registry is the Radiation Dose Structured Report (RDSR) and localizer. We accept secondary capture (or dose screen) because, at the time the Dose Index Registry (DIR) was created, there were many old scanners that were not able to produce RDSRs and we wanted to accommodate those scanners. The ACR uses optical character recognition (OCR) software to read the data that comes from secondary captures. After facilities begin transmitting data to the registry, they can log into their account and confirm that the data have been received by the DIR for your facility. Facilities can review the volume of exams received and are able to check which scanners are successfully submitting data. The ACR encourages monthly data quality checks on the scanners to ensure accurate data flow.
- **Issue 5:** The measure developer should provide results by the levels of specification - clinician group and facility - not aggregated results.
    - **Developer response 5:** Per NQF Testing Requirement, although score-level testing of the composite measure score is desired, at initial endorsement empirical or face validity testing of the components OR face validity of the composite is acceptable. At this point, we have not yet conducted empirical validity testing for this measure. However, we have provided additional face validity data in Validity Developer Response #1 above.
- **Issue 6:** The developer provided minimal information with regard to stratification. The measure should be stratified by patient size, so each stratum is compared to size-based DRLs.
    - **Developer response 6:** Stratification by patient-size is built into the measure definition. The measure assesses whether each record has a DLP below a size specific benchmark that applies to the patient associated with that record.
    - The radiation technique required for images of diagnostic quality of any body part varies by patient size; hence our benchmarks vary by patient size. As part of the paper that developed the size-specific diagnostic reference levels underlying the measure, we conducted regression analysis to assess significance of difference in dose indices by patient size. On the basis of this analysis, the final size bins were constructed from a clinical perspective and for practical usefulness. Size bins were 2cm wide for head exams, and 4cm wide for abdomen-pelvis and chest, based on the number of data points in each bin.

Table 4. Comparison of DLP by patient size bins

	N	Mean	StdErr	Median	P25	P75
CT Abdomen-pelvis With Contrast						
<25	1071843	412.62	0.38	313.37	213.80	491.27
25 to <29	1707749	472.08	0.25	405.62	288.55	574.52
29 to <33	2278700	657.56	0.27	577.59	422.43	811.80
33 to <37	1806768	894.67	0.37	810.94	591.50	1085.23
37 to <41	1038415	1100.37	0.59	1003.79	739.28	1324.34
41+	1093874	1404.32	0.70	1263.19	949.29	1694.82
CT Chest Without Contrast						
<25	209331	227.31	0.59	164.19	97.48	256.96
25 to <29	632346	239.61	0.28	193.30	123.11	284.67
29 to <33	1117103	300.59	0.24	253.95	147.09	388.80
33 to <37	984000	397.20	0.31	356.78	189.98	522.19
37 to <41	512215	519.07	0.55	470.72	268.58	672.75
41+	413959	644.99	0.77	561.93	358.35	790.04
CT Head Without Contrast						
<14	4210321	849.18	0.28	781.89	568.85	991.17
14 to <16	4257564	988.31	0.28	893.80	725.67	1080.20
16 to <18	2557221	985.34	0.38	878.39	702.71	1077.66
18 to <20	1658580	956.70	0.42	879.88	710.00	1058.29
>20	979341	1163.53	1.07	919.43	716.63	1153.51

- The underlying variable, Dose Length Product, demonstrates variation by patient size categories in **Table 4;** the differences in mean DLP across size bins are found to all be statistically significant using ANOVA ($p<0.001$). We compared the performance rates of facilities on the stratified and un-stratified measure, by patient size. The un-stratified measure disregarded patient size and did not make any allowances for higher dose for large patients or require smaller doses for small patients.
- We present the descriptive statistics and correlations by size in **Table 5**. The stratified and un-stratified measures are highly correlated, but the un-stratified measure overestimates performance at smaller sizes and underestimates performance at larger sizes relative to the stratified measure.

Table 5. Descriptive statistics and correlations by size.

Abdomen-pelvis	Stratified measure				Unstratified measure				Correlation coefficient
Effective diameter	N	Mean	StdErr	Median	N	Mean	StdErr	Median	
<25	5,786	66.06	0.48	83.20	5,786	95.47	0.16	98.80	0.3656
25 to <29	6,382	64.72	0.44	78.00	6,382	91.54	0.22	98.80	0.5924
29 to <33	6,462	63.23	0.44	78.00	6,462	79.43	0.36	93.60	0.8460
33 to <37	6,383	64.80	0.43	78.00	6,383	60.13	0.44	67.60	0.9742

37 to <41	6,091	62.81	0.44	72.80	6,091	41.10	0.44	36.40	0.8060
41+	5,889	63.97	0.43	72.80	5,889	24.18	0.35	15.60	0.6049

Chest	**Stratified measure**				**Unstratified measure**				**Correlation coefficient**
Effective diameter	**N**	**Mean**	**StdErr**	**Median**	**N**	**Mean**	**StdErr**	**Median**	
<25	5,577	76.34	0.41	88.40	5,577	91.51	0.24	98.80	0.6151
25 to <29	6,370	74.32	0.39	88.40	6,370	90.74	0.24	98.80	0.6821
29 to <33	6,475	74.12	0.38	88.40	6,475	83.07	0.32	93.60	0.8879
33 to <37	6,373	77.19	0.36	88.40	6,373	70.53	0.39	83.20	0.9304
37 to <41	6,004	78.45	0.35	88.40	6,004	55.84	0.44	62.40	0.7233
41+	5,612	81.16	0.34	93.60	5,612	42.28	0.44	36.40	0.5330

Head, brain	**Stratified measure**				**Unstratified measure**				**Correlation coefficient**
Lat thickness	**N**	**Mean**	**StdErr**	**Median**	**N**	**Mean**	**StdErr**	**Median**	
<14	5,772	56.06	0.51	67.60	5,772	65.23	0.47	78.00	0.8984
14 to <16	6,371	60.60	0.46	72.80	6,371	66.41	0.44	83.20	0.9435
16 to <18	6,173	63.98	0.46	78.00	6,173	63.21	0.46	78.00	0.9948
18 to <20	4,710	65.20	0.54	83.20	4,710	60.36	0.56	72.80	0.9556
>20	4,427	61.09	0.57	72.80	4,427	49.02	0.59	52.00	0.8562

- The risk stratification analysis is only performed at the level of the facility and not group. This is because groups are generally aggregations of facilities – a group supports one or more facilities. Any findings applicable at the facility level formulation of the measure applies to group level as well.