Scientific Methods Panel Fall 2019 Discussion Guide

Measures for Discussion

Subgroup 1

- 2456 Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient (Brigham and Women's Hospital)
  - Reliability: H-0; M-4; L-2; I-0 Pass, Moderate
  - Validity: H-0; M-3; L-2; I-1 CNR
- 1623 Bereaved Family Survey (Department of Veterans Affairs / Hospice and Palliative Care)
  - Reliability: H-3; M-2; l-0; I-1 Pass, High
  - Validity: H-1; M-1; L-3; I-1 CNR
- 0575 Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Control (<8.0%) (National Committee for Quality Assurance (NCQA))
  - Reliability: H-1; M-4; L-0; I-0 Pass, Moderate
  - Validity: H-2; M-2; L-0; I-1 Pass, Moderate
- 0059 Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Poor Control (>9.0%) (NCQA)
  - Reliability: H-2; M-3; L-0; I-0 Pass, Moderate
  - Validity: H-1; M-3; L-0; I-1 Pass, Moderate
- 0061 Comprehensive Diabetes Care: Blood Pressure Control (<140/90 mm Hg) (NCQA)
  - Reliability: H-2; M-3; L-0; I-0 Pass, Moderate
  - Validity: H-3; M-1; L-0; I-1 Pass, High
- 0425 Functional status change for patients with lumbar impairments (Focus on Therapeutic Outcomes, Inc. (FOTO))
  - Reliability H-3; M-1; L-0; I-1 Pass, High
  - Validity: H-4; M-1; L-0; I-0 Pass, High

Subgroup 2

- 0696 STS CABG Composite Score (The Society of Thoracic Surgeons (STS)
  - Reliability: H-0; M-6; L-1, I-0 Pass, Moderate
  - Validity: H-2; M-1; L-3; I-1 CNR
  - Composite Construction: H-3; M-2; L-1; I-1 Pass, High
- 3537 Intraoperative Hypotension among Non-Emergent Noncardiac Surgical Cases (Mathematica)
  - Reliability: H-4; M-0; L-2; I-1 CNR
  - Validity: H-1; M-3; L-3; I-0 CNR
- 0018 Controlling High Blood Pressure (NCQA)
  - Reliability: H-4; M-1; L-0; I-2 Pass, High
  - Validity H-0; M-4; L-2; I-1 CNR
- 3534 30 Day All-cause Risk Standardized Mortality Odds Ratio following Transcatheter Aortic Valve Replacement (TAVR). (American College of Cardiology)
  - Reliability: H-0; M-2; L-3; I-2 Fail, Low
  - Validity: H-0; M-4; L-2; I-1 CNR

Subgroup 3

- 3478 Surgical Treatment Complications for Localized Prostate Cancer (Alliance of Dedicated Cancer Centers)
  - Reliability: H-0; M-3; L-1; I-1 CNR
  - Validity: H-0; M-3; L-1; I-1 CNR
- 3492 Acute Care Use Due to Opioid Overdose (Yale CORE / CMS)
  - Reliability: H-4; M-0; L-1; I-2 CNR
  - Validity: H-1; M-3; L-0; I-3 CNR

Subgroup 4

- 3528 CDC and VON Harmonized Outcome Measure for Late Onset Sepsis and Meningitis in Very Low Birthweight Neonates (Centers for Disease Control and Prevention (CDC))
  - Reliability: H-1; M-1; L-3; I-1 Fail, Low
  - Validity: H-0; M-4; L-2; I-0 Pass, Moderate
- 3483 Adult Immunization Status (NCQA)
  - Reliability: H-4; M-1; L-0; I-1 Pass, High
  - Validity: H-2; M-3; L-1; I-0 Pass, Moderate
  - Composite: H-5; M-0, L-1, I-0 Pass, High
- 3484 Prenatal Immunization Status (NCQA)
  - Reliability: H-4; M-1; L-0; I-1 Pass, High
  - Validity: H-2; M-3; L-1; I-0 Pass, Moderate
  - Composite: H-5; M-0, L-1, I-0 Pass, High

Measures that Passed (Not Pulled for Discussion) [Appendix A]

Subgroup 1

- 2651 CAHPS® Hospice Survey (experience with care) (Centers for Medicare & Medicaid Services (CMS))
  - Reliability: H-2; M-4; L-0; I-0 Pass, Moderate
  - Validity: H-0; M-6; L-0; I-0 Pass, Moderate
- 3533e Hospital Harm – Severe Hyperglycemia (IMPAQ International LLC)
  - Reliability: H-6; M-0; L-0; I-0 Pass, High
  - Validity: H-4; M-1; L-0; I-1 Pass, High

Subgroup 2

- 0071 Persistence of Beta-Blocker Treatment After a Heart Attack (NCQA)
  - Reliability: H-2; M-5; L-0; I-0 Pass, Moderate
  - Validity: H-0; M-5; L-1; I-1 Pass, Moderate

Subgroup 3

- 3538 All-Cause Emergency Department Utilization Rate for Medicaid Beneficiaries Who May Benefit from Integrated Physical and Behavioral Health Care (CMS, Centers for Medicaid & CHIP Services)
  - Reliability: H-5; M-1; L-0; I-0 Pass, High

- Validity: H-2; M-4; L-0; I-0 Pass, Moderate
- 0684 Percent of Residents with a Urinary Tract Infection (long stay) (CMS)
  - Reliability: H-0; M-5; L-1; I-0 Pass, Moderate
  - Validity: H-1; M-3; L-1; I-1 Pass, Moderate

Subgroup 4

- 2979 Standardized Transfusion Ratio for Dialysis Facilities (CMS)
  - Reliability: H-2; M-3; L-1; I-0 Pass, Moderate
  - Validity: H-4; M-0; L-0; I-2 Pass, High
- 3543 Patient-Centered Contraceptive Counseling (PCCC) measure (UCSF)
  - Reliability: H-5; M-1; L-0; I-0 Pass, High
  - Validity: H-5; M-1; L-0; I-0 Pass, High

Subgroup 1

Measure# 2456: Number of Unintentional Medication Discrepancies per Patient

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure assesses the actual quality of the medication reconciliation process by identifying errors in admission and discharge medication orders due to problems with the medication reconciliation process. The target population is any hospitalized adult patient. The time frame is the hospitalization period. At the time of admission, the admission orders are compared to the preadmission medication list (PAML) compiled by trained pharmacist (i.e., the gold standard) to look for discrepancies and identify which discrepancies were unintentional using brief medical record review. This process is repeated at the time of discharge where the discharge medication list is compared to the PAML and medications ordered during the hospitalization.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Data, Electronic Health Records, Instrument-Based Data, Other, Paper Medical Records
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: "The patient denominator includes a random sample of all potential adults admitted to the hospital. Our recommendation is that 25 patients are sampled per month, or approximately 1 patient per weekday."
  - **Ratings for reliability:** 4 moderate and 2 low → Measure passes with MODERATE rating
  - Reliability testing conducted at the data element level:
    - Developer tested reliability of the data elements with an inter-rater reliability assessment, wherein two study pharmacists independently collected medication histories for 19 randomly-selected patients, calculating the percentage of patients for whom there was complete agreement in medication, dose, route, and frequency across the two assessments.
      - Results: 77% agreement

        - In addition, the developers evaluated inter-rater reliability of the discrepancy scoring system by analyzing the last 4 quarterly cases, consisting of a total of 44 medications and 128 ratings each for admission and discharge discrepancies.
        - Results:
            - For the presence of admission discrepancies, the developer found agreement for 116/128 ratings (91% agreement)
                - Kappa = 0.64 (substantial agreement)
            - For the presence of discharge discrepancies, the developer found agreement for 116/128 ratings (91% agreement)
                - Kappa = 0.64 (substantial agreement)
    - **Ratings for validity:** 3 moderate, 2 low, and 1 insufficient → Consensus not reached
        - To demonstrate validity, the developer notes that the literature shows that pharmacists take more accurate medication histories than either nurses or physicians, suggesting that a preadmission medication history taken by a trained expert pharmacist is itself a reasonable proxy for a "gold standard" medication history. The developer's implication is that the measure is using the gold standard itself to identify discrepancies in medication histories.
        - The developers provided materials to show how expert pharmacists are trained, as well as materials showing that the process used to measure discrepancies is transparent and systematic.
        - In addition, the developers provide some data showing that 9 of 17 study sites had significant improvement in their discrepancy rates in the last 6 months of the study compared to the first 6 months of the study. Those that did improve rates had a greater increase in the number of patients receiving recommended patient-level interventions.

ITEMS TO BE DISCUSSED

- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
- RELIABILITY:
    - Should there be a minimum number of patients required to calculate this measure?
    - If sampling is involved, the method should be specified.
    - One Panel member suggests the relevant Standing Committee should review the training and instruction materials to make sure they are adequate/show that a "gold standard" history is actually being collected – the measure is only as good as this process
    - Availability of trained pharmacists for all sites raises scalability concerns
    - Denominator statement is unclear
    - Some reviewers expressed concerns about the reliability testing
        - Result of 77% agreement (without correcting for chance) is not high, calls into question subsequent evaluation results
        - Testing should be based on more cases – 4 not enough

        - Kappa of .64 is “low but potentially acceptable”; “moderate but still pretty good”
- VALIDITY
    - One reviewer suggested that excluding patients who are unavailable to be seen by a pharmacist or who decline to talk to the pharmacist could introduce bias (e.g., cultural, SES) into the results. Is there any alternative to live interviews to get medication histories?
    - One reviewer questioned the exclusion of “patients who are discharged before a gold standard medication list can be obtained” as this could potentially be used as an excuse for not getting data on difficult patients.
        - **Action item:**
            - To address this, should a time limit be set (e.g., those discharged in 6 hours or less)?
    - Validity testing
        - Reviewers noted that while the measure appears to have face validity from a common sense perspective, face validity was not systematically assessed, and suggested that there was a lack of data provided to support the developer’s assertion that the measure can reliably identify discrepancies, even if it uses the gold standard for gathering medication histories.
            - **Action items:**
                - Is the developer’s rationale for face validity of the measure adequate?
                - Is use of the ‘gold standard’ practice for collecting medication histories sufficient to demonstrate validity?
        - The developers provide data suggesting improvement in measure performance may be associated with more consistent implementation of best practices for gathering medication histories.
            - **Action Item:**
                - Does this provide additional support for the measure’s validity?

Measure# 1623: Bereaved Family Survey

MEASURE HIGHLIGHTS

- Maintenance Measure, previously reviewed and did not pass SMP evaluation
- **Description:** This measure calculates the proportion of Veteran decedent´s family members who rate overall satisfaction with the Veteran decedent´s end-of-life care in an inpatient setting as "Excellent" versus "Very good", "good", "fair", or "poor".

  The measure is derived from item #18 of the 20-item Bereaved Family Survey (English and Spanish versions): “*Overall, how would you rate the care that [the Veteran decedent] received in the last month of life*?”

  The measure is calculated as the number of respondents who choose “Excellent” (v. all other responses) divided by the number of completed BFS [defined as surveys with a valid response for item 18 plus at least 12 more valid responses on the forced-choice items).
- **Type of measure:** Outcome: PRO-PM

- **Data source:** Instrument-Based Data [Bereaved Family Survey]
- **Level of analysis:** Facility
- **Risk-adjusted (using statistical risk model with 5 factors)**
- **NQF responses on PA comments**
    - MIF document: items 1b,3, and 4. These sections of the submission form are not due to NQF until November 1 (note that they address other evaluation criteria, not Scientific Acceptability)
    - "Waste" of respondent data: Many PROMs include various domains that may be aggregated into performance measures (PRO-PMs). However, NQF does not require that all instrument items "feed into" PRO-PMs. Moreover, each PRO-PM is evaluated as a separate measure by NQF, even if several are included under one NQF number.
    - Previous evaluation by SMP: NQF was not clear regarding the previous evaluation conducted by the SMP in Fall 2018. Specifically, the testing attachment for that evaluation was provided to the SMP, but without sufficient explanation (it was provided only for informational purposes). It seems possible that at least one SMP member looked at this testing attachment rather than the most current one when rating the measure. The testing attachment that should be considered for this current evaluation is named "**BFS_TestingAttachment_073019**".
- **Ratings for reliability:** 3 high, 2 moderate, and 1 insufficient → Measure passes with HIGH rating
    - To demonstrate reliability of the survey item used in this measure, the developers conducted 4 test-retest analyses on 93 randomly selected BFS respondents who agreed to complete the BFS on a 2nd occasion (30 days apart)
        - Analysis #1 (Cohen's kappa): Kappa=0.5 (n=92); Developer cites Cohen's article that says a kappa of 0.5 indicates moderate agreement
        - Analysis #2: two-way random effects, absolute agreement, single rater/measurement: ICC (2,1)=0.52 (moderate agreement, according to Cohen)
        - Analysis #3, Logistic Regression: Compared to those who reported BFS=0 at time 1, respondents who reported BFS=1 at time 1 had 17.2 the odds of reporting BFS=1 at time 2 (interpreted as very strong association)
        - Analysis #4, Cohen's d Effect Size of a 2x2 contingency table: d=1.57 ("large" effect when d≥0.8)
    - Developers also described an analysis of the global item obtained via phone vs. mail administration (2009-2012 data for phone, 2012-2017 data for mail). Results indicate both are normally distributed (mean=58, 63 respectively, both with SD=5), and very few facilities had mean ≥ 90, interpreted as no ceiling effects. Developers also reported Cronbach's alpha for phone vs. mail (0.81 vs 0.83). NOTE that it is not clear whether these analyses meet NQF's requirements for data element reliability testing of the global item (NOTE that the developers provide additional clarification about their methods in their responses). If not, ratings for data element reliability should be based on the test-retest analysis described above.
    - To demonstrate reliability of the measure score, the developers conducted 2 analysis:
        - ICC1 using a mixed-effects logistic regression model:
            - FY10-FY12 (administered predominantly as a phone survey)

        - Facility-level variance estimate=0.15; 95% CI .12-.20; p<0.001
        - ICC1=0.04 (95% CI: .03-.06)
      - FY13-FY17 (administered predominantly via a mail survey):
        - Facility-level variance estimate =0.13; 95% CI .09-.20; p<0.001.
        - ICC1=0.04 (95% CI: .03-.06)
    - Split-half analysis with application of Spearman-Brown prophecy formula: 0.89
    - NOTE: The panel member who rated as insufficient may not have seen the test-retest results
- **Ratings for validity:** 1 high, 1 moderate, 3 low, and 1 insufficient → Measure does not pass
  - To demonstrate validity of the survey item used in this measure, the developers analyzed 5% (randomly selected) of written responses to the question *"Is there anything else that you would like to share about the Veteran's care during the last month of life?"* These comments were categorized as positive, neutral, or negative. These categorizations were correlated with the responses from the overall rating of care item (the item from the survey used in this measure).
    - Spearman correlation coefficient=0.51; p<0.001
  - To demonstrate validity of the measure score, the developers compared the results of this measure to those obtained from 4 process measures using 4 linear regressions adjusted for nonresponse bias and patient case-mix (n=146 facilities). The outcome was this measure (% with "excellent"), and the independent variable was the process measure. Their hypothesis was that receipt of each of the "best practices" processes should result in a statistically significant higher BFS score.
    - Palliative Care Consult prior to death
      - Phone survey: β=0.03; 95% CI (0.03-0.03), p-value<.001
      - Mail survey: β=0.03; 95% CI (0.02-0.03), p-value<.001
    - Death in a Hospice/Palliative Care Unit
      - Phone survey: β=0.03; 95% CI (0.03-0.03), p-value<.001
      - Mail survey: β=0.02; 95% CI (0.02-0.03), p-value<.001
    - Chaplain Contact with Veteran or Family
      - Phone survey: β=0.02; 95% CI (0.02-0.03), p-value<.001
      - Mail survey: β=0.03; 95% CI (0.02-0.04), p-value<.001
    - Bereavement Contact with Family
      - Phone survey: β=0.01; 95% CI (0.01-0.02), p-value<.001
      - Mail survey: β=0.02; 95% CI (0.01-0.02), p-value<.001
    - The results of these analyses support the developer's hypotheses.
      - In the analysis of exclusions/missing data, a total of 16% of eligible decedent veterans were excluded from the measure. A total of 4% were excluded because they died within 24 hours of admission. The remaining excluded cases were included in a nonresponse bias analysis.
      - The measure is risk-adjusted, using 5 factors (veteran's age at the time of death; number of medical comorbidities present at the time of death; veteran's primary diagnosis on last admission; relationship of veteran's next-of-kin (i.e., spouse), and model of administration mode (i.e., mail).

ITEMS TO BE DISCUSSED

- **This measure did not pass on reliability in the Fall 2018 evaluation by the SMP** due to lack of reliability testing at the data element level (i.e., for item #18 in the BFS). There was also a question of whether the score-level validation was done with the risk-adjusted measure. The developers have subsequently conducted test-retest reliability analyses in order to meet NQF's requirements for data element testing.
- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
- VALIDITY
    - Testing
        - At least one panel member questioned whether score-level testing was done with the risk-adjusted score. Page 10 on the testing attachment indicates that the developers DID test using the risk-adjusted score. Also see developer response, p.3.
        - At least one panel member may have rated validity only based on an analysis of trends in scores across time. However, this analysis was not included as part of the validity testing in the most recent testing attachment for the measure. Therefore, this panel member may not have used the most current information when rating the measure.
        - Concern that the statistically significant associations of this measure with the four process measures are due to large sample sizes. Also, the β coefficients are quite low. There was also some question of whether a multivariate model would perform better.
            - In their response, the developers note that construct validation does not require a recommended magnitude of association. They emphasize that their analysis demonstrates "theoretically-predictable, positive, and significant" correlations.
            - The question of how a multivariate model might perform MAY have arisen if the panelist was not aware that the analyses were conducted using case-mix adjusted scores.
        - **Action Items:**
            - Discuss the score-level testing and results presented in the current testing attachment.
                - Do concerns remain regarding testing using the risk-adjusted score?
                - Consider the low associations found in the testing results. Should these low values be interpreted as absence of construct validation of the measure?
            - Discuss the information provided to demonstrate validity of the survey item (i.e., correlation with comments). Does this meet NQF's requirements for data element validation of the survey item?
    - Exclusions
        - The earlier testing attachment incorrectly indicated the lack of exclusions to the measure. This was fixed in the current testing attachment.

        - There were some questions about excluding veterans with no family members or with no contact information on family members (i.e., does this introduce selection bias), as well as the exclusion for surveys where <12 items are answered.
            - In their additional responses (p.7), the developers note that <1% of surveys have fewer than 12 items answered.
        - One SMP member desired clarification regarding nonresponse and "low-complexity" facilities and the potential effect on validity of the measure.
            - In their additional responses, the developers suggest these usually are VA nursing homes. They say this is included in the risk-adjustment approach. NOTE that this does not align with the factors listed by the developers, ***so this MUST be clarified***.
        - **Action Item:** Discuss the exclusions for the measure. Do they seem warranted?
    - Risk Adjustment
        - SMP members noted a lack of clarity regarding what is included in the risk-adjustment approach, why those variables were considered, why some were ultimately not included.
            - In their additional responses, the developers discuss rationale for inclusion of comorbidities (p.6) and their rationale for not including social risk variables (p.6).
        - NOTE that NQF encourages SMP members to note concerns about inclusion/lack of inclusion of risk-factors but does not allow this to be a reason for "failing" a measure.
        - **Action Item:** Discuss the risk-adjustment approach. Do you understand what factors are included and why (including how nonresponse and facility complexity is handled)? Do you agree with the rationale provided by the developer?

Measure# 0575: Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Control (<8.0%) (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) whose most recent HbA1c level is <8.0% during the measurement year.
    - Numerator: Patients whose most recent HbA1c level is less than 8.0% during the measurement year.
    - Denominator: Patients 18-75 years of age by the end of measurement year who had a diagnosis of diabetes (type 1 and type 2) during the measurement year or the year prior to the measurement year.
    - Exclusions
        - This measure excludes adults in hospice.
        - It also excludes adults with advanced illness and frailty, as well as Medicare adults 65 years of age and older enrolled in an I-SNP or living long-term in institutional settings.

- Additionally, exclude patients who had a diagnosis of gestational diabetes or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year and who did NOT have a diagnosis of diabetes.
    - These patients are sometimes pulled into the denominator via pharmacy data.
    - They are then removed once no additional diagnosis of diabetes (Type I or Type II) is found.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records
- **Level of analysis:** Health Plan
- **Not risk-adjusted**
- **Ratings for reliability:** 1 high and 4 moderate → Measure passes with MODERATE rating
    - Score level testing was conducted using the beta-binomial methodology defined by Adams
    - 401 commercial plans, 250 Medicaid plans, and 477 Medicare plans were analyzed
    - Table 2. Overall Beta-binomial statistic and distribution of plan reliability for commercial, Medicaid, and Medicare product lines, 2018

Product Line	Overall Reliability	Min	Percentiles					Max
			10th	25th	50th	75th	90th	
Commercial	0.995	0.808	0.978	0.978	0.979	0.983	0.995	1.000
Medicaid	0.978	0.611	0.885	0.949	0.952	0.957	0.961	1.000
Medicare	0.975	0.768	0.964	0.968	0.969	0.976	0.979	1.000

- **Ratings for validity:** 2 high, 2 moderate, and 1 insufficient → Measure passes with MODERATE rating
    - Score level testing was conducted using correlation analyses for construct validity
    - Developer tested for construct validity of the Comprehensive Diabetes Care (CDC): HbA1c Control (<8.0%) measure by exploring whether it was correlated with other similar measures of quality hypothesized which are listed below.
        - CDC: Hemoglobin A1c (HbA1c) Testing: The percentage of adults 18-75 with diabetes that had an HbA1c test performed during the measurement year.
        - CDC: HbA1c Poor Control (> 9.0%): The percentage of adults 18-75 with diabetes whose most recent HbA1c level is >9% during the measurement year.
        - CDC: Eye Exam (Retinal) Performed: The percentage of adults 18-75 with diabetes that had an eye screening for diabetic retinal disease during the measurement year.

        - CDC: Medical Attention for Nephropathy: The percentage of adults 18-75 with diabetes that had a nephropathy screening test or evidence of nephropathy during the measurement year.
        - CDC: Blood Pressure Control (<140/90 mm Hg): The percentage of adults 18-75 with diabetes whose most recent blood pressure level taken during the measurement year is <140/90 mm Hg.
    - Results ranged from 0.35 to 0.99 indicating moderate to very strong correlation.

ITEMS TO BE DISCUSSED

- RELIABILITY
    - Concerns raised that the beta-binomial approach and the reliability score obtained with this approach may not support the assertion that "the higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another." If sample is large enough, it can be easy to achieve a very high reliability score. (Testing form page 5, 2a2.2).
        - Beta-binomial: Discuss the "overall reliability" meaning from beta-binomial results.
    - Concerns raised that the systematic sampling method described in S.15 is prone to bias.
- VALIDITY
    - Concerns regarding the clarify of the t-test described in 2b4.1, in that it appears to be just comparing two proportions.
    - Concerns with the lack of clarity regarding the treatment of missing.
    - Concerns that the exclusion rate for advanced illness and frailty criteria to not accurately reflect prevalence for known conditions (e.g., heart failure). (Testing form page 10, 2b2.2)
    - Concerns raised that testing to demonstrate comparable results for multiple data sources was not provided.

Measure# 0059: Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Poor Control (>9.0%) (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) whose most recent HbA1c level is >9.0% during the measurement year.
    - Numerator: Patients whose most recent HbA1c level is greater than 9.0% or is missing a result, or for whom an HbA1c test was not done during the measurement year.
    - Denominator: Patients 18-75 years of age by the end of measurement year who had a diagnosis of diabetes (type 1 and type 2) during the measurement year or the year prior to the measurement year.
        - There are two ways to identify patients with diabetes: by claim/encounter data and by pharmacy data. The organization must use both methods to identify the eligible population, but a patient only needs to be identified by one method to be included in the measure. Patients may be identified as having diabetes during the measurement year or the year prior to the measurement year.

- This measure excludes adults in hospice.
- It also excludes adults with advanced illness and frailty, as well as Medicare adults 65 years of age and older enrolled in an I-SNP or living long-term in institutional settings.
- Additionally, exclude patients who had a diagnosis of gestational diabetes or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year and who did NOT have a diagnosis of diabetes. These patients are sometimes pulled into the denominator via pharmacy data. They are then removed once no additional diagnosis of diabetes (Type I or Type II) is found.

- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records
- **Level of analysis:** Health Plan
- **Not risk-adjusted**
- **Ratings for reliability:** 2 high, 3 moderate → Measure passes with MODERATE rating
  - Score level reliability was tested using the beta-binomial approach.
  - Developer tested data from 378 commercial plans, 241 Medicaid plans, and 477 Medicare plans:

Product Line	Overall Reliability	Min	Percentiles					Max
			10th	25th	50th	75th	90th	
Commercial	0.996	0.831	0.980	0.982	0.983	0.987	0.992	1.000
Medicaid	0.983	0.627	0.915	0.955	0.966	0.973	0.977	1.000
Medicare	0.980	0.792	0.970	0.975	0.977	0.982	0.985	1.000

- **Ratings for validity:** H-1; M-3; L-0; I-1 → Measure passes with MODERATE rating
  - Developer tested for construct validity by Pearson's correlations with other similar measures of quality hypothesized which are listed below:
    - CDC: Hemoglobin A1c (HbA1c) Testing: The percentage of adults 18-75 with diabetes that had an HbA1c test performed during the measurement year.
    - CDC: HbA1c Control (<8.0%): The percentage of adults 18-75 with diabetes whose most recent HbA1c level is <8% during the measurement year.
    - CDC: Eye Exam (Retinal) Performed: The percentage of adults 18-75 with diabetes that had an eye screening for diabetic retinal disease during the measurement year.
    - CDC: Medical Attention for Nephropathy: The percentage of adults 18-75 with diabetes that had a nephropathy screening test or evidence of nephropathy during the measurement year.
    - CDC: Blood Pressure Control (<140/90 mm Hg): The percentage of adults 18-75 with diabetes whose most recent blood pressure level taken during the measurement year is <140/90 mm Hg.
  - These measures exhibited moderate to very strong inverse correlation, ranging from -0.32 to -0.99.

- Note: The correlation values are all negative because the HbA1c Poor Control measure is a "lower is better quality" measure, while the other measures are "higher is better". This indicates that plans that have low rates on this measure will have high rates on the others.

ITEMS TO BE DISCUSSED

- RELIABILITY
  - The meaning of "overall reliability" included in Table 2 was questioned by at least one panel member and the accuracy of the results.
  - Concern of systematic bias was raised by several reviewers. There are multiple data sources listed; should data element level testing be conducted to ensure consistent results?
  - One panel member asked the question: should between vs. within plan variation be assessed with intraclass correlation coefficients with plotted plan measures and standard errors?
- VALIDITY
  - Concerns that the exclusion rate for advanced illness and frailty criteria to not accurately reflect prevalence for known conditions (e.g., heart failure). (Testing form page 10, 2b2.2)

Measure# 0061: Comprehensive Diabetes Care: Blood Pressure Control (<140/90 mm Hg) (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) whose most recent blood pressure level taken during the measurement year is <140/90 mm Hg.
  - Numerator: Patients whose most recent blood pressure level was <140/90 mm Hg during the measurement year.
  - Denominator: Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 and type 2) during the measurement year or the year prior to the measurement year.
    - This measure excludes adults in hospice.
    - It also excludes adults with advanced illness and frailty, as well as Medicare adults 65 years of age and older enrolled in an I-SNP or living long-term in institutional settings.
    - Additionally, exclude patients who had a diagnosis of gestational diabetes or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year and who did NOT have a diagnosis of diabetes. These patients are sometimes pulled into the denominator via pharmacy data. They are then removed once no additional diagnosis of diabetes (Type 1 or Type II) is found.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records
- **Level of analysis:** Health Plan

- **Not risk-adjusted**
- **Ratings for reliability:** 2 high and three moderate → Measure passes with MODERATE rating
    - Reviewers expressed concerns with the lack of clarity of the specifications (e.g., how compliance is determined, exclusions)
    - Testing of performance measure score with beta binomial reliability.
    - Tests conducted on 394 commercial plans, 250 Medicaid plans and 477 Medicare plans; Overall reliability ranged from 0.976 to 0.998 respectively.
- **Ratings for validity:** H-3; M-1; L-0; I-1 → Measure passes with HIGH rating
    - To establish construct validity, the developer correlated this measure to other measures that are similarly focused on diabetic patients.
        - CDC: Hemoglobin A1c (HbA1c) Testing: The percentage of adults 18-75 with diabetes that had an HbA1c test performed during the measurement year.
        - CDC: HbA1c Control (<8.0%): The percentage of adults 18-75 with diabetes whose most recent HbA1c level is <8.0% during the measurement year.
        - CDC: HbA1c Poor Control (> 9.0%): The percentage of adults 18-75 with diabetes whose most recent HbA1c level is >9% during the measurement year.
        - CDC: Eye Exam (Retinal) Performed: The percentage of adults 18-75 with diabetes that had an eye screening for diabetic retinal disease during the measurement year.
    - Correlation scores ranged from 0.41 to 0.89, indicating moderate to strong correlations.
    - Reviewers expressed concern regarding lack of analysis on use of multiple data sources and comparability of results when more than one source was available for a plan; lack of clarity around the handling of missing data.
    - Some reviewers expressed concern regarding the lack of risk adjustment for clinical factors and the developers rationale for this decision.

ITEMS TO BE DISCUSSED

- RELIABILITY
    - Specifications:
        - There was some confusion as per whether a patient would be counted as compliant if both systolic and diastolic BP values are in compliance or if a patient would be counted as compliant if systolic and diastolic BP values are in compliance.
        - The appropriateness of the telehealth exclusion was questioned, especially from the rural lens.
    - Beta-binomial results: One panelist questioned whether it is likely that the within plan variation is low enough to yield a reliability coefficient >.90, and suggested it would have been useful to see the plan level ICCs and a distribution of plan level rates with standard error bars. Discuss the need for this.

- VALIDITY

- Discuss the concern that multiple data sources are used but comparisons of results between them are not made.
- Discuss the developer's approach to dealing with missing data.
- Discuss the need for risk adjustment in this measure. One panelist was not convinced that there should not be a robust consideration of risk factors for this measure.

Measure# 0425: Functional Status Change for Patients with Low Back Impairments (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** A self-report outcome measure of functional status for patients 14 years+ with lumbar impairments. The change in functional status assessed using FOTO (lumbar) PROM is adjusted to patient characteristics known to be associated with functional status outcomes (risk adjusted) and used as a performance measure at the patient level, at the individual clinician, and at the clinic level by to assess quality.
  - Numerator: The numerator is based on residual scores (actual change scores - predicted change after risk adjustment) of patients receiving care for Low Back impairments and who completed the Low Back PRO-PM. The numerator, as it applies to the 3 levels, is defined as follows:
    - Patient Level: The residual functional status score for the individual patient with a low back impairment.
    - Individual Clinician Level: The average of residuals in functional status scores in patients who were treated by a clinician in a 12-month time period for a low back impairment.
    - Clinic Level: The average of residuals in functional status scores in patients who were treated by a clinic in a 12-month time period for a low back impairment.
  - Denominator: The target population is all patients 14 years and older with a Low Back impairment who have initiated an episode of care and completed the Low Back FS PROM.
    - FOTO recommends that clinicians have a minimum of 10 patients/year and clinics have a minimum of 10 patients/therapist per year for small clinics or 40 patients per year for larger clinics (5 or more clinicians) in order to obtain stable estimates of provider performance.
    - Exclusions: Patients who are not being treated for a Low Back impairment. Patients who are less than 14 years of age.
- **Type of measure:** Outcome: PRO-PM
  - Residuals are calculated based on baseline and final patient reports
- **Data source:** Instrument-Based Data
- **Level of analysis:** Clinician: Group/Practice, Clinician: Individual
- **Risk Adjustment:** Adjusted (Statistical Risk Model)
- **Ratings for reliability:** 3 high, one moderate, and 1 insufficient → Measure passes with HIGH rating
  - Reliability testing was conducted at the data element and score level
    - Data element testing was conducted through tests of internal consistency of the instrument, and reliability of point estimates and change scores

        - Score level was evaluated using the beta binomial methodology defined by Adams.
- Internal consistency test had a person reliability estimate of 0.92
    - Reliability of point estimates and change scores: on average, the mean of 95% CI upper limits of MDC95 values for all patients was 13.9, but the mean MDC95 value for 97% of patients with FS initial evaluation scores between 20 and 80 was 7.8.
    - Beta binomial calculations:

Reliability (R) at the provider level: 2016-2017								
	Number of patients with complete episodes per clinician per calendar year	Variance explained (%) by the provider level	N providers	Average R	Min R	Max R	N if R≥0.7	% if R≥0.7
Clinic	*FOTO	5.8	3098	0.84	0.21	1.00	2674	86
	20+	5.8	2942	0.86	0.41	1.00	2636	90
	30+	5.5	2732	0.87	0.48	1.00	2523	92
	40+	5.5	2520	0.88	0.55	1.00	2397	95
Clinician	10+	6.8	12025	0.71	0.19	0.98	7029	58
	20+	6.8	7787	0.77	0.37	0.98	5618	72
	30+	6.9	4849	0.81	0.50	0.98	4191	86
	40+	7.4	2867	0.84	0.57	0.98	2799	98
*10+ per clinician for small clinics (1-3 clinicians), 40+ per clinic for large clinics (4 or more clinicians) Acceptable levels of reliability are marked in green								

- **Ratings for validity:** H-4; M-1; L-0; I-0 → Measure passes with HIGH rating
    - Validity testing was conducted at the data element and score level
    - Data element testing was conducted via content validity tests, structural validity, local independence and item fit, differential item functioning, construct validity, sensitivity to change, responsiveness and content range coverage, and clinically important improvement.
    - Score level testing was done by assessing performance score level by MCII achievement, and correlation of performance scores with two external markers (GROC and ODQ)
    - Data element results:
        - After removing 3 items with low factor loadings and/or poor fit, the 25-item pool represented a unidimensional pool with strong local independence.
        - After removing three items, confirmatory factor analysis results for the remaining 25 items were CFI = 0.87, TLI = 0.98, and RMSEA = 0.09 for the one-factor solution, demonstrating a unidimensional item pool with strong local independence.

        - Only the BPFS items describing working and driving displayed nonuniform DIF by gender and age, respectively (P<0.002). No items displayed uniform DIF. DIF adjusted and unadjusted Low Back FS ability estimates were highly correlated (i.e., r values all >0.9992).
        - Sensitivity to change: Results support that the Low Back FS PROM was sensitive to change. The initial evaluation FS measures averaged 51 (SD=12), discharge FS scores were 65 (SD=16), and FS change scores were 14 (SD=16), which produces an effect size ([discharge minus initial evaluation]/[standard deviation of initial evaluation]) of 14/12=1.17, which is considered large.
        - Responsiveness: Results suggested that the Low Back FS PROM was responsive, and MCII was dependent on initial evaluation FS with patients perceiving improvement with fewer FS units as initial evaluation FS scores increase.
    - Score level results:
        - A higher proportion of patient episodes managed by higher performing providers experienced change equal to or greater than the MCII as compared to lower performing providers. This pattern was observed using both methods of provider performance ranking; uncertainty assessments (3 levels) and percentile ranking (10 levels).
        - Validity of clinician performance based on uncertainty assessments (3 levels):
            - The three performance levels had statistically significant differences between groups as determined by one-way ANOVA ($F(2,12022) = 4342.7$, $p < 0.001$) with a monotonic increase in rates of MCII achievement.
            - For Low Back PRO-PM correlations with the GROC, a sample of 202 clinics and 924 clinicians were included. Low Back PRO-PM correlations with the ODQ, a sample of 208 clinics and 669 clinicians were included. Absolute correlations for the two measures and provider levels ranged from 0.62 to 0.78 (see TABLE 2b1.3ix below) and were highly significant (P<0.001).

ITEMS TO BE DISCUSSED

- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
- RELIABILITY
    - Reviewers expressed concerns with the inconsistencies in the specifications (e.g., age range for inclusion). Clarify the age restrictions of the measure: 8 vs. 14 years as cutoff points.
    - Concerns about whether this measure reaches adequate levels of reliability the for assessing the performance of clinicians with 10-19 and 20-29 cases.

- VALIDITY

- Concerns with "Structural validity" evidence and clarity of the patient-level discriminant analysis. Construct validity evidence appears to be missing.
- Lack of clarity in presentation of data (e.g., the number of clinicians per clinic and the potential impact on confounding of clinician with clinic performance; data regarding sample sizes in Table 2b1.3ix do not correspond with data in Table 2b1.3vi)
- Concerns raised that more testing on SES factors should be done. There is no discussion of the sources of or causes for the time between onset of symptoms and initial evaluation, but to the extent it is due to lack of access to providers, there may be a related SES component.
- Concerns raised with approach for assessing education as a risk factor.
- Concerns with whether testing adequately addresses whether patients excluded resulted in bias in measurement. (e.g., those completing vs not on some patient characteristics, there is no assessment on social risk factors and the impact on completion rates, an alternate completing the survey, clinician completing survey). Questions on whether this should have been included in the risk adjustment model as another variable, or at least tested for significance.

Subgroup 2

Measure# 0696: STS CABG Composite Score

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The STS CABG Composite Score comprises four domains consisting of 11 individually NQF-endorsed cardiac surgery measures:
  - Domain 1) Absence of Operative Mortality – Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death during the same hospitalization as surgery or after discharge but within 30 days of the procedure;
  - Domain 2) Absence of Major Morbidity – Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as having at least one of the following adverse outcomes (all or none to enter numerator):
    - 1. reoperations for any cardiac reason,
    - 2. renal failure,
    - 3. deep sternal wound infection,
    - 4. prolonged ventilation/intubation,
    - 5. cerebrovascular accident/permanent stroke;
  - Domain 3) Use of Internal Mammary Artery (IMA) – Proportion of first-time CABG patients who receive at least one IMA graft;
  - Domain 4) Use of All Evidence-based Perioperative Medications – Proportion of patients who receive all required perioperative medications for which they are eligible. The required perioperative medications are (all or none to enter numerator):
    - 1. preoperative beta blockade therapy,
    - 2. discharge anti-platelet medication,
    - 3. discharge beta blockade therapy,
    - 4. discharge anti-lipid medication.

All measures are based on audited clinical data collected in a prospective registry. Participants receive a score for each of the domains, plus an overall composite score. The overall composite score is created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by one star (below average performance), two stars (average performance), or three stars (above average performance). Scores and star ratings are currently publicly reported on STS and Consumer Reports websites.

- **Specifications:** weighing—81% mortality, 10% morbidity, 7% IMA, 3% medications
- **Type of measure:** Composite, maintenance, rate/proportion where higher is better and star rating as well 1=below average…3=above average. Weights are applied.
- **Data source:** Registry
- **Level of analysis:** Group/Practice
- **Exclusion:** Those with contraindications internal mammary artery use, or of various drug use (e.g., contraindications of beta blockers). About 3% of cases excluded because of these reasons. Exclusions also made based on small sample sizes.
- **Risk-adjusted:** Yes, morbidity and mortality measures adjusted with 48 risk factors as was done previously for measure 0119. Variables: age, body surface, surgical date, ejection fraction, creatinine, sex, and dialysis. Calibration of model was solid with C-stat= 0.75 and observed versus expected plots highly correspondent.
- **Ratings for reliability:** 6 moderate and 1 low → Measure passes with MODERATE rating
    - Performance level signal to noise (SNR) with Bayesian approach to calculation true probability for the reliability testing. SNR= 0.68 with considerable spread.
- **Ratings for validity:** H-2; M-1; L-3; I-1 → **Consensus not reached**
    - Measure gap: recent cohort of 1,024 practices (2014) showed performance of .97 (SD= .00092) across 143K procedures (presumed weighted and adjusted).
    - Pearson's correlation between two time periods July 2013-2014 vs. July 2012-2013 was 0.64; Spearman's 0.63.
    - Face validity results suggested, but not described in great detailed.
    - Meaningful differences somewhat evident, even as about 80% perform at average levels (see table below).

Table X. Star ratings in the last four harvests

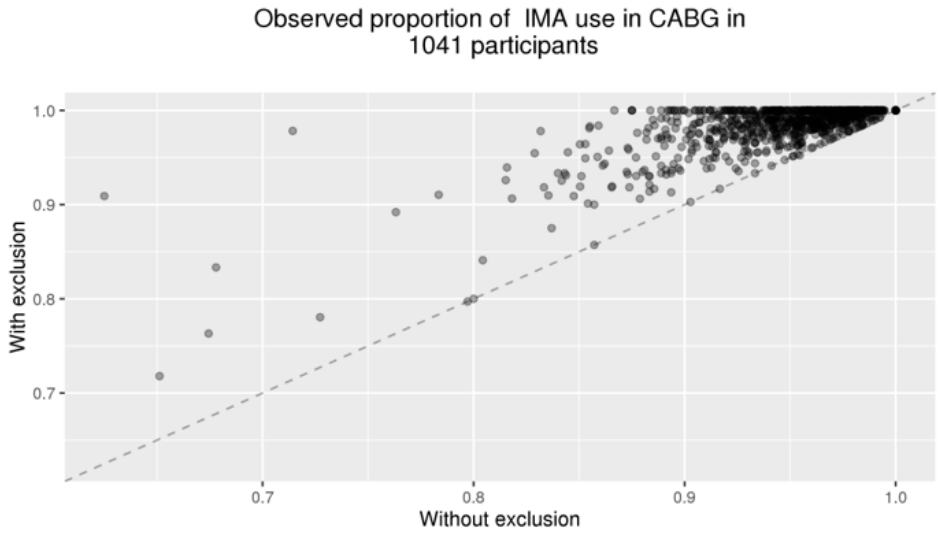
Star Rating	07/2013-06/2014	01/2013-12/2013	07/2012-06/2013	01/2012-12/2012
1	60, 5.9%	98, 9.6%	97, 9.6%	91, 9.0%
2	864, 84.4%	770, 75.7%	782, 77.3%	770, 76.5%
3	100, 9.8%	149, 14.7%	132, 13.1%	146, 14.5%

- Correlations between star rating and domain score also presented, but arguably a bit circular in logic.
- **Ratings for Composite Construction:** H-3; M-2; L-1; I-1 → Passes with HIGH Rating
    - 11 measures that lie beneath all are NQF endorsed
    - Compilation method that all must be achieved, for numerator of composite to be fulfilled
    - Weighted construction: see note under specifications

- Missing data: rare 1/1,000 observations, and only 13 of centers had more the 5% missing data, and results did not change.

ITEMS TO BE DISCUSSED

- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
- VALIDITY
  - **Action Item:** From 2012 to 2014? See **Table X** above which shows trends that suggests that as of this year (2019) most providers may be concentrated in the 2 Star cluster.
    - Discuss whether meaningful differences can still be detected today and whether there may be concerns with the measures being topped-out?
      - In response to this critique, the developer requested more time to update their testing. See Appendix B for additional information submitted by developer.
  - Exclusions (e.g., removing those contraindicated for certain drugs or the mammary graft) appear to markedly increase performance rates (compared to absence of exclusion).
    - **Action Item:** Consider whether this is this a threat to validity. See figure below.



The Spearman rank correlation of the measures with and without the exclusion is 0.60. The correlation is 0.73.

  - **Action Item:** Is the weighting scheme for the adjusted performance score properly justified? 81% mortality, 10% morbidity, 7% IMA, 3% medications
  - **Action Item**: Discuss the correlations between Star rating and domain score.
  - **Action Item:** There were concerns that no external standard was used to demonstrate validity. The developer provided a response:

        - "With most individual measures, it is possible to find another external quality metric against which to assess validity. In the case of the STS CABG composite, we have purposely included all the major quality metrics that have been used for CABG. Thus, they are within the composite and are not available as separate external measures for validation. That is the reason we showed the correlation of the overall composite score with results for each of the domains."
    - **Action Item:**
        - Risk Adjustment: Discuss the backward selection approach for the risk adjustment model that resulted in variable errors of omission.

Measure# 3537: Intraoperative Hypotension among Non-Emergent Noncardiac Surgical Cases

MEASURE HIGHLIGHTS

- New measure
- **Description:** Percentage of noncardiac, non-emergency surgery cases involving general anesthesia or monitored anesthesia care (MAC) of adults (ages 18 and older) in which mean arterial pressure (MAP) fell below 65 mmHg for cumulative total of 15 minutes or more.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Registry
- **Level of analysis:** Individual Clinician
- **Risk-adjusted** (using a logistic regression model with five risk variables)
- **Ratings for reliability:** 4 high, 2 low, and 1 insufficient → Consensus not reached
    - Conducted signal-to-noise analysis using Adams method
    - Developer provides the distribution of reliability coefficients by clinician denominator size and by quartile
    - Reported reliability coefficients ranged from .87 to .98
- **Ratings for validity:** 1 high, 3 moderate, and 3 low → Consensus not reached
    - Developer tested the measure using two forms of measure score validity testing: predictive validity and known-group validity.
    - *Predictive validity* of the measure was assessed by examining the association between the unadjusted measure score and the negative downstream outcomes linked to intraoperative hypotension (IOH) in the clinical literature.
        - Patients who experienced negative downstream outcomes—AKI, MINS, or in-hospital mortality—had significantly higher rates of IOH than patients who did not experience these outcomes, as expected.
    - *Known-group validity* was assessed by testing whether anesthesia cases involving patient sub-populations known to be at greater risk for IOH had significantly poorer measure scores than anesthesia cases not involving those high-risk sub-populations.
        - Consistent with the literature, patients with a higher ASA physical status classification and patients with a lower BMI had

significantly higher rates of IOH than patients with lower ASA and higher BMI respectively.

- Contrary to the literature, however, patients age 65 or older had slightly but significantly lower IOH rates than patients under age 65.

ITEMS TO BE DISCUSSED

- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
  - This additional information includes analyses intended to provide further support for measure reliability and validity:
    - Reliability coefficients reported by case volume, showing wider variation in reliability scores (see pp. 6-7):

Distribution of reliability coefficients for risk-adjusted IOH measure in 2016, by number of denominator cases

	Clinicians with 1-30 cases	Clinicians with 31-100 cases	Clinicians with 101-200 cases	Clinicians with 201-500 cases	Clinicians with 501+ cases
Number of clinicians	78	87	78	216	207
Percent of sample (N = 666 clinicians)	12%	13%	12%	32%	31%
Mean	0.50	0.83	0.94	0.97	0.98
5th ptile	0.04	0.71	0.90	0.95	0.97
10th	0.11	0.73	0.91	0.95	0.98
20th	0.20	0.77	0.93	0.96	0.98
30th	0.26	0.79	0.93	0.96	0.98
40th	0.37	0.82	0.93	0.97	0.98
50th	0.53	0.84	0.94	0.97	0.99
60th	0.58	0.86	0.95	0.97	0.99
70th	0.66	0.87	0.95	0.98	0.99
80th	0.71	0.89	0.97	0.98	0.99
90th	0.74	0.91	0.97	0.99	0.99

Distribution of reliability coefficients for risk-adjusted IOH measure in 2017, by number of denominator cases

	Clinicians with 1-30 cases	Clinicians with 31-100 cases	Clinicians with 101-200 cases	Clinicians with 201-500 cases	Clinicians with 501+ cases
Number of	88	85	102	208	215

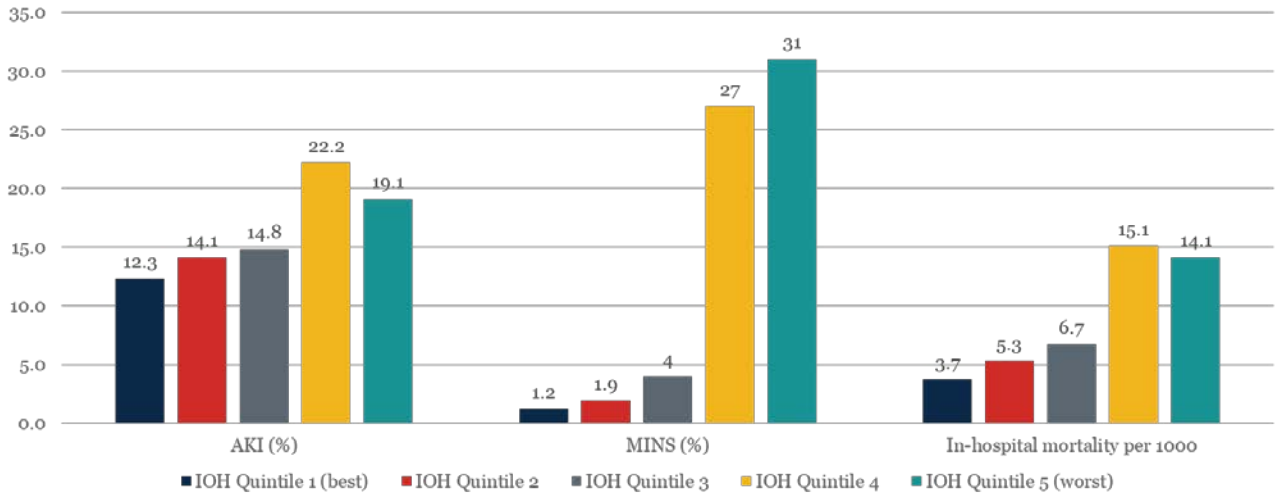
	Clinicians with 1-30 cases	Clinicians with 31-100 cases	Clinicians with 101-200 cases	Clinicians with 201-500 cases	Clinicians with 501+ cases
clinicians					
Percent of sample (N = 698 clinicians)	13%	12%	15%	30%	31%
Mean	0.39	0.81	0.92	0.96	0.98
5th ptile	0.07	0.67	0.85	0.93	0.97
10th	0.09	0.69	0.88	0.94	0.97
20th	0.14	0.73	0.90	0.95	0.97
30th	0.25	0.79	0.91	0.96	0.98
40th	0.36	0.81	0.91	0.96	0.98
50th	0.43	0.83	0.92	0.96	0.98
60th	0.48	0.84	0.93	0.97	0.98
70th	0.54	0.85	0.93	0.97	0.99
80th	0.61	0.87	0.96	0.98	0.99
90th	0.66	0.89	0.97	0.98	0.99

- Incidence of adverse outcomes by clinician-level, risk-adjusted IOH score quintile (see pp. 19-20):

Developer Response 23: The new graph below provides evidence of the validity of the risk-adjusted, clinician-level IOH score. We calculated the rates of adverse patient outcomes (acute kidney injury [AKI], myocardial injury after noncardiac surgery [MINS], in-hospital mortality), by risk-adjusted IOH score quintile. This analysis was done in Site 1 only, as Site 2 did not provide data on AKI or MINS.

The graph shows that clinicians with the worst 40 percent of risk-adjusted scores (quintiles 4–5) have meaningfully higher rates of AKI, MINS, and in-hospital mortality compared with clinicians with the best 60 percent of scores (quintiles 1–3). The relationship is particularly strong for MINS and in-hospital mortality.

Incidence of AKI, MINS, and in-hospital mortality, by clinician-level, risk-adjusted IOH score quintile



- Do the updated analyses adequately substantiate measure reliability and validity?

Measure #0018: Controlling High Blood Pressure (NCQA)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The percentage of adults 18-85 years of age who had a diagnosis of hypertension (HTN) and whose blood pressure was adequately controlled (<140/90 mm Hg) during the measurement year.
- **Type of measure:** Intermediate Outcome
- **Data source:** Claims, Electronic Health Data, Electronic Health Records, Paper Medial Records
- **Level of analysis:** Health Plan
- **No Risk-adjustment:** No analysis of social factors completed; used plan type as proxy for income and socioeconomic status.
- **Ratings for reliability:** 4 high, 1 moderate, 2 insufficient → Measure passes with HIGH rating
  - Reliability of the health plan measure score was tested using a beta-binomial approach (i.e., signal to noise); Overall reliability ranged 0.982-0.999 across the three types of plans
  - Reviewers expressed concerns with clarity and consistency of specifications (i.e., different age ranges used throughout the specifications, lack of clarity around target blood pressure, inconsistencies in the denominator and numerator details)
- **Ratings for validity:** 4 moderate, 2 low, 1 insufficient → Consensus was not reached; must be discussed by subgroup
  - Validity of the health plan measure was demonstrated through construct validity using the entire HEDIS data sample.
    - Construct validity of the Controlling Blood Pressure measure was conducted by assessing the correlation with another measure: Comprehensive Diabetes Care: Blood Pressure Control (<140/90 mm Hg): The percentage of adults 18-75 years

of age with diabetes (type 1 and type 2) whose most recent blood pressure level taken during the measurement year is <140/90 mm Hg.
    - Pearson correlation across the three types of health plans ranged 0.75 to 0.93; Medicare with the lowest and commercial plans with the highest correlation score.
    - Concerns with validity included lack of analysis around multiple data sources, analysis of exclusions, lack of risk adjustment, and the measure used to correlate for construct validity.

ITEMS TO BE DISCUSSED

VALIDITY

- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
- Construct Validity Testing:
    - Reviewers expressed concern with the measure selected to test correlation and demonstrate construct validity. The measure selected also measures blood pressure control, but for a diabetic population. There is concern that there may be some overlap in the denominator populations and data elements between the measures. Further, it was unclear to some reviewers whether the high correlation of these two measures demonstrated empirical/predictive validity of the measure in the absence of additional information.
        - Does the measure selected to demonstrate construct validity represent an independent measure of quality?
- Risk adjustment:
    - This measure is not risk adjusted. Reviewers expressed concern with the lack of risk adjustment and whether the stratification of plan types is sufficient to account for the differences in the populations. In addition to socioeconomic differences, there may be a need to adjust for clinical factors that may make it more difficult for some patients to achieve the target populations than others.
        - Is patient risk adequately accounted for in the measure's current construction and specification (i.e., exclusions and stratification)?
- Multiple data sources
    - The measure is specified for four different data sources including, claims, electronic health data, electronic health records, and paper medical records. There is currently no analysis included in the submission on how the data collected from these data sources might impact reliability or validity.
        - How might data collected from these various data sources impact reliability and/or validity? What additional analyses should the developer consider to analyze any impact these data sources may have on the measure results?
- Exclusions
    - The developer performed an analysis of the exclusions and determined that on average about 3.8% of the sample is excluded based on the measure's specifications. However, there was no sensitivity analysis that would validate that exclusions were appropriately applied.

- Is the exclusion analysis provided adequate? If not, what additional analysis should be performed to better support the methodology and validity of the exclusions as specified?

Feedback to developer: Any advice for the developers on how to improve the submission for this cycle or a future cycle?

- Reliability:
  - Clarify the number of minimum cases required to achieve reliability
  - Provide sample size per plan type/product line
  - Consider how data type and data collection method might impact reliability
- Validity:
  - To improve the demonstration of construct validity, identify an independent measure to make an empirical association between the implicit quality construct and the material outcome (or better yet an explicity quality construct and the material outcome). For example, that health plans with worse performance on controlling high blood pressure among hypertension / diabetes patients have worse performance on coronary artery disease, congestive heart failure, stroke, ruptured aortic aneurysm, renal disease and retinopathy.

Measure# 3534: 30 Day All-cause Risk Standardized Mortality Odds Ratio following Transcatheter Aortic Valve Replacement (TAVR).

MEASURE HIGHLIGHTS

- New measure
- **Description:** This measure estimates hospital risk standardized 30-day all-cause mortality odds ratios following transcatheter aortic valve replacement. The measure uses clinical data available in the STS/ACC TVT Registry for risk adjustment. For the purpose of development and testing, the measure used site-reported 30-day follow-up data in the STS/ACC TVT Registry.

  The Risk Standardized Odds Ratio is calculated as the odds that an outcome (e.g. 30-day mortality) will occur for patients treated at a facility compared to the "odds" that outcome will occur for patients with identical risk factors if treated by a hypothetical (average) hospital. Thus, a lower odds ratio implies lower-than-expected mortality (better quality) and a higher ratio implies higher-than-expected mortality (worse quality).
- **Type of measure:** Outcome
- **Data source:** Registry Data [STS/ACC TVT Registry]
- **Level of analysis:** Facility
- **Risk-adjusted (**using statistical risk model with 41 factors)
- **Ratings for reliability:** 2 moderate, 3 low, and 2 insufficient → Measure does not pass
  - To demonstrate reliability of the data elements used in the measure, the developer assessed inter-rater reliability using data from 40 records selected randomly from 4 randomly selected facilities (presumably, 10 records per facility, although this is not clear)
- **Ratings for validity:** 4 moderate, 2 low, and 1 insufficient → Consensus not reached
  - To demonstrate validity of the data elements, the developers conducted 2 analyses:

        - Record eligibility assessment: 6 hospitals participating in the registry reported all TAVT and Mitral cases performed at their facility during a specified timeframe. These were compared to those records included in the registry to verify that cases were not missed. N=366 records
        - 40 hospitals with at least 10 cases were randomly selected for audit. From each hospital, 10 baseline and 10 follow-up cases (for 30-day and 1-year) were randomly selected for abstraction. Sample included 400 "baseline" records, 400 "30-day" records, and 289 "1-year" records. Developers calculated the prevalence-adjusted and bias-adjusted kappa (PABAK) statistic.
    - Key concerns in the SMP's initial analysis regarding the measure include exclusion of >50% of hospital/patients due to missing data, relatively low values of PABAK for two tested values, lack of data element testing for most variables, and a relatively small testing sample that may or may not be representative of hospitals/patients included in the measure.

ITEMS TO BE DISCUSSED

- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
    - These included inter-rater reliability estimates for several additional variables (these data reflect different data years; however, it is unclear if the methodology was the same as that presented in the original submission materials).
- RELIABILITY
    - Testing
        - To demonstrate reliability of the data elements used in the measure, the developer assessed inter-rater reliability using data from 40 records selected randomly from 4 randomly selected facilities (presumably, 10 records per facility, although this is not clear). Data for this analysis was abstracted by two trained auditors.
            - Developers note the IRR assessment was performed on baseline, and 30-day and one-year follow-up cases, although it is not completely clear what this means.
            - This analysis found a 100% agreement rate for the five variables tested (discharge status, discharge date, follow-up status, five meter walk test performed, and KCCQ-12 performed). They had no data on the follow-up date of death variable.
            - NQF requires analysis beyond percent agreement (e.g., ICC, kappa). However, the developers may not have done this testing due to the 100% agreement found for 5 of the tested variables. This should be clarified.
        - SMP concerns with the data element reliability testing included:
            - Only 5 data elements used for the measure were tested

        - Per NQF requirements, testing at the level of data elements requires that all critical data elements be tested. At a minimum, the numerator, denominator, and exclusions (or exceptions) must be assessed and reported separately. When defining critical data elements, NQF states that "*testing at the data element level should include those elements that contribute most to the computed measure score, that is, account for identifying the greatest proportion of the target condition, event, or outcome being measured (numerator); the target population (denominator); population excluded (exclusions); and when applicable, risk factors with largest contribution to variability in outcome*."
    - Only 40 records were used in testing, and no information was provided about the 40 records (or 4 facilities) included in the sample.
        - Per NQF requirements, testing may be conducted on a sample of the accountable entities. The sample should represent the variety of entities whose performance will be measured. Ideally, all types of entities whose performance will be measured should be included in reliability and validity testing. The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method. When possible, units of measurement and patients within units should be randomly selected.
    - Abstractions from trained auditors was used in the testing. However, it not clear how these trained auditors compare to abstractors in the field who presumably send data to the registry.
    - It is not clear if any of the patients represented in the 40 sample records died in the 30-day period.
- **NOTE** that if data element validation is conducted (and results are deemed adequate), NQF does not require additional reliability testing. Thus, if the reliability testing as described is not considered sufficient, the SMP should also consider any data element validation that was conducted and apply these results to their ratings of reliability.
- **Action Items:**
    - Is the number of hospitals/patients included in testing sufficient to determine reliability?
    - Is the lack of ICC or kappa statistics reasonable, given the percent agreement found in testing?
    - Should other variables be considered "critical" data elements and therefore be used in testing? ***NOTE: See additional materials (pp1-3)***
    - If the reliability testing is not considered adequate, will the data element validity testing that was presented suffice for a moderate rating for reliability?

- NOTE that because score-level reliability testing was not conducted, the highest eligible rating for reliability is MODERATE.

- VALIDITY
  - Testing
    - Record eligibility assessment: 6 reported all TAVT and Mitral cases performed at their facility during a specified timeframe (n=366 records). These were compared to those records included in the registry to verify that cases were not missed.
      - 365 of the 366 cases were included in the registry (one was omitted)
    - 40 hospitals with at least 10 cases were randomly selected for audit. From each hospital, 10 baseline and 10 follow-up cases (for 30-day and 1-year) were randomly selected for abstraction. Sample included 400 "baseline" records, 400 "30-day" records, and 289 "1-year" records.
      - Discharge status: PABAK=1.0
      - Discharge date: PABAK=0.97
      - Follow-up status: PABAK=0.77
      - Follow-up date of death: PABAK=0.50
      - Five-meter walk test performed: PABAK=0.82
      - KCCQ-12 performed: PABAK=0.92
    - No score-level validation was conducted, although the developers referred to calibration plots examined as part of their risk-adjustment analyses.
      - NQF does not accept calibration statistics, alone, as score-level validity testing.
    - SMP concerns with the data element validity testing included:
      - Only 6 data elements used for the measure were tested
      - Low PABAK values for follow-up date of death variable (0.50) and follow-up status (0.77)
    - **Action items**:
      - The submission materials did not clarify which variable(s) reflect the outcome of interest (i.e., death within 30 days). Has this been clarified sufficiently?
      - Is there any concern regarding the relatively low PABAK values for follow-up date of death and follow-up status?
      - Should other variables be considered "critical" data elements and therefore be used in testing? ***NOTE: See additional materials (pp6-8)***
  - Risk-adjustment
    - This measure is risk-adjusted via hierarchical logistic regression mode with site-specific random intercept parameters (41 factors included)
      - C-statistic: 0.703
      - Calibration was assessed via risk-decile plots (overall and for pre-specified subgroups (age, sex, ejection fraction, NYHA class, prior aortic procedure))
    - Race, ethnicity, and Medicaid status were considered for inclusion in the risk-adjustment approach. Seemingly, none of these were associated with 30-day

mortality. Participation in Medicaid was NOT included as a factor in the risk-adjustment model. Race/ethnicity WAS included in the final risk-adjustment model. However, the developers note they are not viewed, conceptually, as an indicator of socio-economic status and state they were included as a way to enhance face validity of the adjustment approach and to reduce confounding.
        - NOTE that NQF encourages SMP members to note concerns about inclusion/lack of inclusion of risk-factors but does not allow this to be a reason for "failing" a measure.
        - **Action items**:
            - Are there any concerns about the risk-adjustment approach?
    - Meaningful differences in performance
        - Some SMP members would have preferred analysis that included confidence intervals.
    - Exclusions/Missing data
        - Exclusion analysis was conducted primarily on the 2016 data that were used to initially develop the measure. The developers note that since that time, data from more hospitals and patients have been included in the registry.
        - Even so, some SMP members expressed concerns:
            - Even though participation in the registry has increased, many hospitals are still not included in the measure.
            - The rationale for the exclusions (i.e., baseline KCCQ-12 score, and baseline gait speed) was not provided.
            - There was no information provided about other exclusions, including how often 30-day mortality is missing.
        - SMP members reiterated concern about the number of hospitals/patients excluded from the measure (more than half, apparently in large part due to missing values for 30-day mortality, KCCQ-12 score, and baseline gait speed), and suggest that the "gain" in the c-statistic/AIC of including the KCCQ-12 score and gait speed is not worthwhile.
            - NOTE that assessing the clinical "value" of including the KCCQ-12 score and gait speed is outside the purview of the SMP.
        - **Action items**:
            - Discuss the statistical implications of excluding patients where 30-day mortality, KCCQ-12 score, and baseline gait speed is missing/not imputed.

Subgroup 3

Measure# 3478: Surgical Treatment Complications for Localized Prostate Cancer

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This measure analyzes hospital/facility-level variation in patient-relevant outcomes during the year after prostate-directed surgery. Specifically, the measure uses claims to identify urinary incontinence and/or erectile dysfunction among patients undergoing localized prostate

cancer surgery (comparing each patient's own claims pre- and post-surgery) and uses this information to derive hospital-specific rates. Those outcomes are rescaled to a 0-100 scale, with 0=worst and 100=best.

- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Ratings for reliability:** 3 moderate, 1 low, 1 insufficient → Consensus not reached
    - Score-level reliability was demonstrated using split half and Pearson's correlations; At the minimum sample size (N=10; 233 hospitals), correlation was ~0.65; the largest sample size (N=80; 28 hospitals) correlation was ~0.89 (*Note: Score OR data element reliability testing is sufficient to meet NQF's requirements)
    - Reviewers expressed concerns with lack of clarity in the specifications (i.e., attribution specifications, timeframes, scoring methodology, and winsorization method)
    - Data set tested: 2009-2013 using ICD-9 codes (*NQF provided a waiver to the developer, which allowed submission of this measure tested with ICD-9 codes; this waiver was granted due to data limitations and developers will be required to submit updated testing for maintenance of endorsement if endorsed)
- **Ratings for validity:** 3 moderate, 1 low, 1 insufficient → Consensus not reached
    - Validity was demonstrated describing a systematic assessment of face validity. This assessment meets NQF requirements for face validity and is acceptable for testing validity of new measure submissions.
    - Some reviewers expressed concern about the lack of risk adjustment, the analysis of exclusions, and missing data.

ITEMS TO BE DISCUSSED

- RELIABILITY
    - Specifications: (developer clarification needed)
        - Concerns raised regarding consistency and clarity of specifications:
            - Denominator eligibility a year prior to the index admission or 1 year after? Settings?
            - Ensure all exclusions are also included in the measure information from (e.g., patients who did not have continuous enrollment in Parts A/B or enrolled in HMO during timeframe)
            - Clarification on how duration of diagnosis is determined using claims data.
            - Clarification on the process used to transform the raw differences (truncated at -5 and 10) to the 0-100 scale. Concerns that this process may present challenges in magnifying differences.
            - How the attribution logic addresses cases when patients have procedures done at more than one facility.
    - Concerns with the reliability and counting of the number of days a patient has a numerator event and the bias potentially introduced by procedures for incontinence or medication for erectile dysfunction.

- VALIDITY
    - Risk Adjustment
        - Concerns raised that even with the analysis to demonstrate that risk adjustment had little impact on the measure score, there may be other conditions that would have in impact on incontinence (e.g., stroke, Parkinson's, diabetes).
            - How might the measure outcome be impacted with the lack of risk adjustment and accounting for the impact of relevant co-occurring conditions?
        - There was acknowledgement by the developers that complications could have an impact, but was unable to obtain data to account for this.
            - How does this potentially impact the measure score and is it feasible to address this concern (given limitations in data)?
    - Missing Data: Clarify how missing data/codes is handled for minimally invasive/robotic procedures. (Developer clarification needed)
        - Consider how missing data impacts reliability for the linking process with SEER. Can the developer clarify how that process was done, whether patient level data was missing and how it was handled?
    - Meaningful Differences: Concerns raised regarding the interpretation of meaningful differences based on the scale provided.
    - Testing: Some reviewers expressed concern with the last of data element validity testing. NQF does not require data element validity testing but only that the developer submits data element OR measure score validity testing. In this case measure score validity is demonstrated with their description of a systematic assessment of face validity.

Measure# 3492: Acute Care Use Due to Opioid Overdose

MEASURE HIGHLIGHTS

- New measure
- **Description:** Late onset sepsis (LOS) is one of the most common complications of extreme prematurity. Studies have indicated that 36% of extremely low gestational age (22-28 weeks) infants develop LOS and that 21% of very low birth weight (VLBW) infants surviving beyond 3 days of life will develop LOS. This infection is usually serious, causing a prolongation of hospital stay, increased cost, and risk of morbidity and mortality. This measure aims to provide summary data.

  Some cases of LOS can be prevented through proper central line insertion and maintenance practices. These are addressed in the CDC's Healthcare Infection Control Practices Advisory Committee (CDC/HICPAC) Guidelines for the Prevention of Intravascular Catheter-Related Infections, 2011. However, almost one-third of LOS events in a quality-improvement study were not related to central-lines.3 Prevention strategies for these non-central line –related infection events have yet to be fully defined, but include adherence to hand-hygiene, parent and visitor education, and optimum nursery design features. Other areas that likely influence the development of LOS include early oral nutritional support and skin care practices.

It is envisioned the use of this measure will promote late-onset sepsis and meningitis prevention activities which will lead to improved patient outcomes including reduction of avoidable medical costs, and patient morbidity and mortality.

- **Numerator:** Number of LOS or Meningitis (MEN) events (for the risk measures), and the survival measure is those without LOS or Men
- **Denominator:** total number of eligible neonates (DOL 4 to 121, and weights/ gestational ages in the range of 400-1500 gms, gestation out of the range of 22-29 weeks, and those in "other than Level II/III or IV nurseries…")
- **Type of measure:** Outcome
- **Data source:** abstracted from electronic health record
- **Level of analysis:** Facility
- **Risk adjustment:** Standardized Infection Ratio (SIR): a ratio of observed-to-predicted infection, across a number of stratified groups. SIR is not calculated when predicted infections are below 0.2.
- **Specification:** 3114 codes for different infection types. Inclusion of definition for a qualified antimicrobial day also given, though not clear why, perhaps as an alternative method to identify an infection, assuming it is not prophylactic.
    - SIR predicated on observed/expected rates and Bayesian statistical methods.
- **Data:** 716 Vermont Hospitals and over 40,000 neonates who met inclusion between 1/10 and 12/16.
- **Ratings for reliability:** 1 high, 1 moderate, 3 low, and 1 insufficient → Measure does not pass
    - Reliability testing not conducted. Instead validity testing was done here using "online calculator" for the numerator only. As such only one data element was tested (numerator) for validity, and thus the measure fails on reliability.
    - Results: "The calculated precision was 100% and recall was 96%. Cohen's kappa coefficient was 0.96, as indicated in the table below. The calculator returned six false negatives due to data entry issues. The hospital chart abstractor at one hospital mis-recorded the names of several bacterial pathogens included on NHSN's list of reportable pathogens, which resulted in false negative calculator determinations." This does not seem to show elements other than infections, per se.
    - 320 cases were selected for precision, and Cohen's kappa:

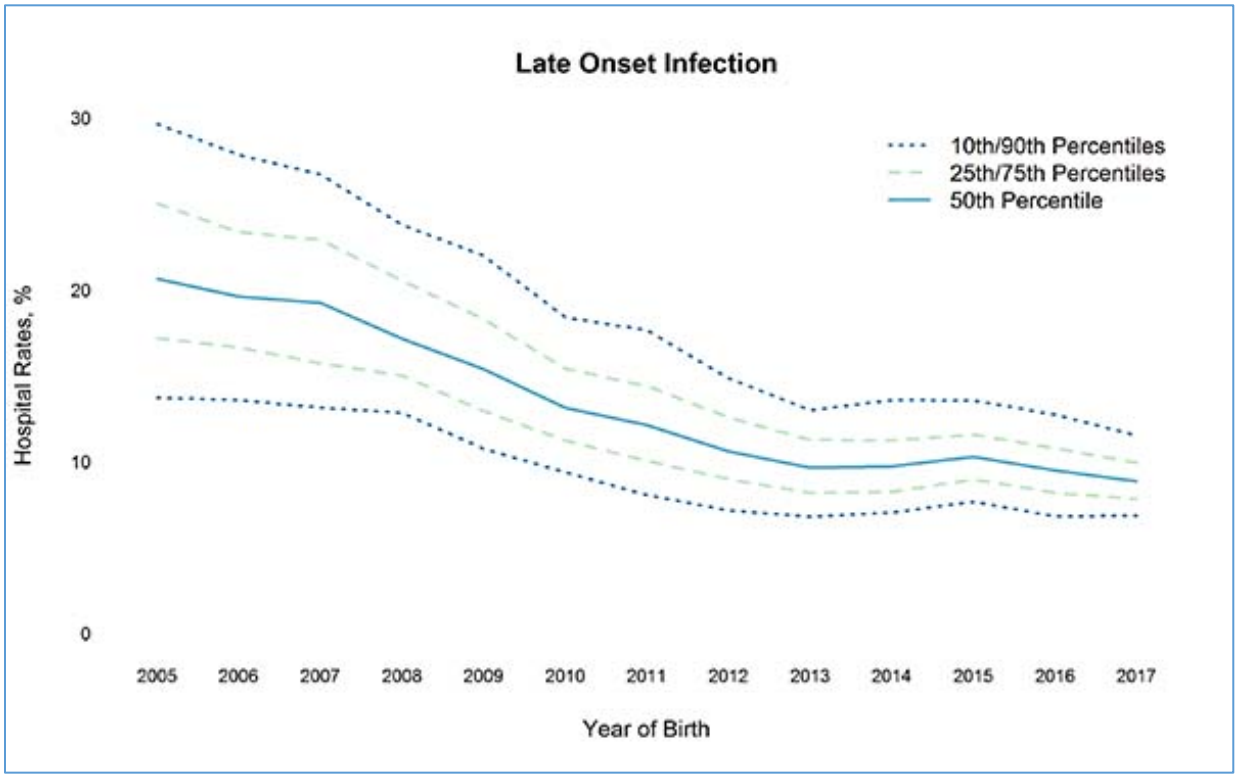
Events		**Manual Abstract**		
		Y	N	*Totals*
Calculator	Y	134	0	134
	N	6	180	186
	Totals	140	180	**320**
	Precision (TP/TP+FP)		100%	
	Recall (TP/TP+FN)		96%	
	Cohen's kappa		0.96	

- **Ratings for validity:** H-0; M-4; L-2; I-0 → Measures passes with MODERATE rating

- Set up not clear, but said to element level
- "Accuracy of data elements by comparing manually abstracted data to gold standard"
- Results: Precision-=.99, Recall=.92, Cohen's Kappa=.94, n=300 cases

- **Risk adjustment:** (section 2b3) Maximum Likelihood estimator deployed for univariate and multivariate comparisons that include the effects of birthweight, gestational age, gender, and whether or not the baby is transferred to a new facility as a neonate.

- **Meaningful differences:** demonstrated via this plot, which actually comes from another measure that is purported to be analogous: NQF 304. Interquartile probabilities differences show here are said to be around 3-4%,which comports with 2017 data show below. Is that enough to suggest a redressable gap?



Missing data: 2b5. Said to be less than 1% missing data, not a problem. No actual analysis presented.

ITEMS TO BE DISCUSSED

- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
- RELIABILITY
  - Reliability (reproducibility) is apparently confused with validity (truth). Here is a comment from the developer about this (underline added for emphasis by NQF staff):
  - **"**The online calculator is an executable, algorithm-based, decision-making tool designed to enable automated determinations of whether individual patients (anonymized) meet the NHSN

surveillance protocol criteria for reportable LOS or MEN events….In effect, the calculator serves as a reference implementation for processing a set of data elements and rendering a rules-based decision concerning reportability. Establishing the calculator's reliability by comparing its performance to expert review of candidate LOS and MEN cases is centrally important to assuring the electronic supply chain produces results that are identical or virtually identical to more labor intensive and costly human expert reviews of source data elements. Our method of reliability testing is designed to demonstrate that the measure data elements are repeatable, producing the same results a high proportion of time when assessed in the same population in the same time period, as per the description of reliability that NQF includes in the MIF (2a2)."

- NQF clarification: this is a case where element level validity can obviate the need for element level reliability. As such, the measure can receive a passing rating.
    - One concern that does remain is that the element level testing does not seem to include the case selection criteria (i.e., the denominator selection approach). See the reporting table above, it suggests just that the numerator (infection) sensitivity/specificity is tested in 320 cases. Perhaps the developer can clarify?

- **Action item**: Revote on reliability, given that element level validity can substitute for it.

Subgroup 4

Measure# 3528: CDC and VON Harmonized Outcome Measure for Late Onset Sepsis and Meningitis in Very Low Birthweight Neonates

MEASURE HIGHLIGHTS

- New Measure; in a previous cycle, did not pass Methods Panel evaluation
- **Numerator statement:** The numerator is comprised of incident outcome events, defined as opioid overdoses that result in emergency department use, within the population residing in a specific geography.
- **Denominator statement:** The denominator consists of all enrolled Medicare Fee-For-Service (FFS) beneficiaries with Parts A or B, aged 18 and older residing in a measured geography (either a county or a state) during a one-year period.
- **Exclusions:** None
- **Description:** This is a claims-based measure that captures the rate of emergency department visits for opioid overdose events using ICD diagnosis codes. Events are measured per 1,000 person-years among Medicare beneficiaries greater than 18 years of age residing in the geography being measured. The measure is designed for use at both the county and state levels.
- **Type of measure:** Outcome
- **Data source:** Medicare Claims
- **Level of analysis:** Population: Community, County or City, Population: Regional and State
- **Not risk-adjusted**
- **Ratings for reliability:** 4 high, 1 low, and 2 insufficient → Consensus not reached
    - Measure score testing:
        - Adams R: 0.92-0.99 across 25 states, 0.6-0.99 across Maryland counties
        - Split sample: 0.94 or 0.87, respectively (Pearson r assumed)
- **Ratings for validity:** H-1; M-3; L-0; I-3 → Consensus not reached

- Face validity adjudicated by 5 Yale physicians only
  - Not many details provided, but Likert voting schemes used
- Comparisons to AHRQ and CDC overdose data
  - Main validity test
    - CDC measures is deaths per 100,000 (correlation with measure = 0.74)
      - NEDS comprehensive ED discharges
      - NIS: comprehensive for ED to inpatient transfers
    - AHRQ measure is opioid-related hospitalizations per 1,000 persons (correlation with measure 0.74)
    - Current measure: opioid-related ED use
  - Meaningful differences
    - Sahai, 1996 method: 95% CI estimates calculated
    - GLM and Poisson distributions to look for year effects
    - Results: 12 states below, 3 at, and 10 above average. Some time variation evident but an explanation was not provided by the developers.

ITEMS TO BE DISCUSSED

- The developers have provided **additional clarifying information** to the SMP, after reviewing the SMP's initial comments on their measure.
- RELAIBILITY
  - Specifications
    - Is geography for the denominator determined by residence or overdose treatment venue- specs are unclear?

      In response the developer added these details about the specs:

    - Denominator is person-years with no minimum period of enrollment
    - Geography of consumer is determined by place of residence, not venue of treatment (and numerator events in other jurisdiction are attributed to a person's place of resident)
    - Medicare advantage enrollees are excluded, Duals (for Medicare and Medicaid) are included
    - "We (the developer) considered specifying the measure with ED visits as the denominator, as one reviewer suggested. However, ED visits are dependent on the baseline health of a population, which may vary from place to place, particularly if the age of the beneficiary population differs from place to place or changes over time. We felt that a denominator that captures the population size is more appropriate."

  - Reliability Testing
    - Concern that the reliability presentation was lacking detail about the data that lies beneath.
      - In response the developer sent the following reliability table sent the following:

- A reliability table (2017 data by state with Adams R all ranging from 0.997-0.992 (n=25), and for Maryland county level jurisdiction ranging from 0.995-0.601 (n=24), only one region below 0.7.
- Split sample reliability was also reported directly, yield correlation coefficients at the state and county levels as 0.94 (df=24) and 0.87 (df=23), respectively. Despite these high coefficients, variability between the random samples is evident (see starred* items in table below)

Split-sample reliability (units= ODs per 1,000 person years)

State	State Measure Results Sample 1	State Measure Results sample 2	County	County Results Sample 1	County Results Sample 2
AZ	0.852	0.875	Allegany	2.128	0.871*
CA	1.032	0.993	Anne Arundel	1.598	1.132
FL	1.272	1.234	Baltimore	2.793	2.104
GA	1.173	1.155	Baltimore City	6.040	6.250
IA	0.586	0.548	Calvert	1.144	0.562
IL	1.017	0.973	Caroline	1.614	1.665
IN	1.370	1.389	Carroll	2.310	1.243
KS	1.076	0.971	Cecil	1.539	2.221
KY	1.591	1.601	Charles	1.449	1.333
MD	1.920	1.682	Dorchester	1.130	2.547
ME	0.873	1.302	Frederick	0.587	0.688
MI	1.831	1.852	Garrett	0.686	1.344
MN	1.469	1.284	Harford	1.719	0.938
MO	1.329	1.272	Howard	0.974	0.568*
MT	0.881	0.740	Kent	0.999	0.656
NC	1.213	1.327	Montgomery	0.440	0.529
ND	0.639	0.555	Prince George's	1.103	0.872
NE	0.696	0.500	Queen Anne's	1.485	0.737*
NV	1.334	1.442	Saint Mary's	1.452	1.577
OR	0.979	1.012	Somerset	2.375	1.856
SD	0.600	0.463	Talbot	0.390	0.791*
TN	1.352	1.356	Washington	1.818	1.814
TX	0.866	0.882	Wicomico	2.176	2.155
WI	0.958	1.005	Worcester	1.135	0.720*
WY	1.077	0.946	--	--	--

* highlight some very large differences between the halves.

- VALIDITY
  - Validity Testing

- What are the potential biases of focusing on Medicare FFS (Parts A+B) patients exclusively?
- Regarding meaningful differences: should there have been more information about the methods used, e.g., statistical tests and additional data to assess their claims of meaningful differences across counties and states?
    - In response the developer submitted the follow information which apparently is a reiteration of what they say was in the application:
        - **"…**we evaluated whether entities (counties or states) differed from the mean using a one sample t-test and we considered a p-value of <0.05…" and found "12 states had below average rates, 10 states were above average, and 3 were no different from average. Among counties, 2 were above average and 9 were below average. In this context, "above average" indicates a higher rate of overdose and worse performance."
        - **"…**used a generalized linear model with a Poisson distribution and a population offset. We fit one model per entity with time (year) as the main effect. A p-value <0.05 for year suggested differences in performance within an entity over time.." Results: "19/25 states had a statistically significant change in measure performance from 2017 to 2018. Among counties… 3 out of 24 counties had a statistically significant change in performance between 2017 and 2018."
- Does absence of any risk-adjustment compromise this measure as a between-region healthcare quality measure? Were socioeconomic issues considered as potential adjusters sufficiently?
- Missing data, was it addressed reasonably?
- Would validity correlation analyses with AHRQ and CDC data be more persuasive if higher versus low volume regions were considered separately?
- Regarding validity, was enough information presented in the application to make the inferences credible?

Measure #3483: Adult Immunization Status (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- New Measure: in a previous cycle, did not pass Methods Panel evaluation
- **Description:** Percentage of adults 19 years of age and older who are up-to-date on recommended routine vaccines for influenza, tetanus and diphtheria (Td) or tetanus, diphtheria and acellular pertussis (Tdap), zoster and pneumococcal.
    - 5 individual measures included under this measure number: influenza, Td or Tdap, zoster, pneumococcal, and the composite
    - Measure is stratified by insurance type (commercial, Medicare, and Medicaid)
- **Type of measure:** Composite
    - Measure reflects the percentage of total recommended vaccines received. Thus, it is not an all-or-none composite and not a more "typical" composite that aggregates independent measures
- **Data source:** Claims, Electronic Health Data, Electronic Health Records, Management Data, Other, Registry Data
- **Level of analysis:** Health Plan, Integrated Delivery System

- **Not risk-adjusted**
- **Ratings for reliability:** 4 high, 1 moderate, and 1 insufficient→ Measure passes with HIGH rating
    - Reliability of the measure score was tested using signal to noise analysis (beta binomial model); Reliability statistics across plans in the 50th percentile were 0.999 to 1.0.
- **Ratings for validity:** 2 high, 3 moderate, 1 low→ Measure passes with a MODERATE rating
    - Developers submitted description of face validity testing as well as empirical testing of the performance score and component measures.
        - Face validity description does not meet NQF requirements; assessment of validity should focus on demonstration of empirical validity.
        - NQF requires that, "Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed."
    - Construct validity was demonstrated using Pearson correlation to assess correlation between measures in the composite, and with other immunization status measures.
- **Ratings for Composite Construction:** 5 high and 1 low → Measure passes with a HIGH rating
    - Developers evaluated the measure components using internal consistency analysis (Cronbach's alpha); Cronbach's alpha ranged 0.948-0.961 across the 3 types of plans.
    - Developers also explored an all-or-none alternative construction of the measure and conducted sensitivity analysis to examine the effects of including/excluding the various components of the measure; performance rates among the plans was higher when the "percentage of total recommended vaccines" approach was used versus the all or none approach.

ITEMS TO BE DISCUSSED

- Members expressed concerns deciphering what may have changed in this submission to address concerns addressed in prior evaluation.
- RELIABILITY
    - Specifications
        - Concerns regarding clarity of the specifications (weighting of components, description of units of measurement in numerator and denominator), and continuity of enrollment
    - Reliability testing
        - Ongoing concerns with the "perfect" range of the reliability score and whether calculations were correctly applied. This was also a concern identified in the SMP's review of this measure in prior cycle (only 3 plans per strata used in estimates); previous reviewers requested shape parameters, details of the calculation and an assessment to support this result. Data from ~135 plans was added to the testing data this cycle to help address this concern, but the Developers did not otherwise provide an assessment that supports the high score.
            - **Action Item:** Did the developers adequately describe how the reliability score was achieved?

- Concerns with the level of analysis "integrated delivery system" indicated in the measure information form and whether the testing aligns with that level of analysis. There does not appear to be testing for integrated delivery system and it was unclear whether this was an intentional selection.
    - **Action Item:** Can the developers clarify which level of analysis was intended? If both were intended, does the health plan level testing suffice to demonstrate reliability at the integrated delivery system level?

For standing committee consideration:

- Meaningful differences in performance: While the IQRs are statistically significant (likely influenced by large sample size), practical difference across health plans aren't described. The largest IQRs observed are among Medicare plans. Consider whether these differences are meaningful from a clinical perspective.

Feedback to developer: Any advice for the developers on how to improve the submission for this cycle or a future cycle?

- To better support reliability testing results: One possible way to provide greater assurance about this (if there is lingering concern about it) would be to report a confidence interval around the estimated reliability. It may also be helpful to report a graphical and/or tabular summary of the raw data (e.g. distribution of sample sizes and measure results) and the parameter estimates from the beta-binomial model.
- Provide more detail on how unequal sample sizes per plan in their reliability calculations was handled.

Measure #3484: Prenatal Immunization Status (Pulled by SMP member)

MEASURE HIGHLIGHTS

- New Measure; in a previous cycle did not pass Methods Panel evaluation
- **Description:** Percentage of deliveries in the measurement period in which women received influenza and tetanus, diphtheria toxoids and acellular pertussis (Tdap) vaccinations.
    - 3 individual measures are included under this measure number: influenza, Tdap, and the all-or-none composite
    - Measure is stratified by insurance type (commercial and Medicaid)
- **Type of measure:** Composite (all-or-none)
- **Data source:** Claims, Electronic Health Data, Electronic Health Records, Management Data, Other, Registry Data
    - Need to determine if developers should add "enrollment data"
- **Not risk-adjusted**
- **Level of analysis:** Health Plan, Integrated Delivery System
- **Ratings for reliability:** 4 high, 1 moderate, and 1 insufficient→ Measure passes with HIGH rating
    - Reliability of the measure score was tested using signal to noise analysis (beta binomial)
        - Overall reliability estimates for both Medicaid and commercial plans was in the 0.99 range

    - Ongoing concerns regarding clarity of the specifications, particularly continuity of enrollment (i.e., 28 days prior to delivery) and how new plan enrollees are included (or not) in the measure.
- **Ratings for validity:** 2 high, 3 moderate, 1 low → Passes with a MODERATE rating
    - Construct validity was demonstrated using Pearson correlation to assess correlation between measures in the composite, and with other immunization status measures.
    - Developers conducted score-level validation of the measures and also described an assessment of face validity
- **Ratings for Composite Construction:** 5 high and 1 low→ Passes with high rating
    - Developers evaluated internal consistency of the component measures using Cronbach's alpha); Cronbach's alpha was 0.948 for commercial plans and 0.958 for Medicaid plans
    - Weighting approach for the component measures was informed by technical expert panels; the individual measures are weighted equally.

ITEMS TO BE DISCUSSED

- Testing only submitted at health plan but specified for both health plan and integrated delivery system levels of analysis.
    - **Action Item:** Can the developers clarify which level of analysis was intended? If both were intended, does the health plan level testing suffice to demonstrate reliability at the integrated delivery system level?

Feedback to developer: Any advice for the developers on how to improve the submission for this cycle or a future cycle?

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

Subgroup 1

Measure# 2651: CAHPS® Hospice Survey (experience with care)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The 8 measures submitted here are derived from the CAHPS® Hospice Survey, which is a 47-item standardized questionnaire and data collection methodology. The survey is intended to measure the experiences of hospice patients and their primary caregivers.
  The measures proposed here include the following six multi-item PRO-PMs.
    - Hospice Team Communication (6 items)
    - Getting Timely Care (2 items)
    - Treating Family Member with Respect (2 items)
    - Getting Emotional and Religious Support (3 items)
    - Getting Help for Symptoms (4 items)
    - Getting Hospice Training (5 items) (applies only for decedents who received hospice care at home or in an assisted living facility)

  In addition, there are two other PRO-PMs, also called "global ratings" that are based on single items.
    - Rating of the hospice care (rating between 0 to 10)
    - Willingness to recommend the hospice

  The survey is completed by the primary caregiver of the patient who died while receiving hospice care (hereafter, "decedent"). The primary caregiver is intended to be the family member or friend most knowledgeable about the decedent's hospice care, and is identified through hospice administrative records. Data collection for sampled decedents/caregivers is initiated two months following the month of the decedent's death. The survey is administered via mail only, telephone only, and mixed mode.

  The measures are calculated only for hospices that have at least 30 completed surveys over eight quarters of data collection. Hospices with 50 to 699 survey-eligible decedents/caregivers in the prior year are required to survey all cases (conduct a census). Hospices with 700 or more survey-eligible decedents/caregivers in the prior year are required to survey a minimum sample of 700 using an equiprobable approach (simple random sampling) and may conduct a census, if desired.

  Measure results are calculated based on "top-box scoring" (i.e., those who selected the most or least positive response) For example, for items where with choices of "Never/Sometimes/Usually/Always" and the most positive response is "always", the top box score is the number of respondents who answer "Always."
- **Type of measure:** Outcome: PRO-PM (8 separate PRO-PMs)
- **Data source:** Instrument-Based Data (instrument is Hospice CAHPS survey)
- **Level of analysis:** Facility
- **Risk-adjusted**

- **Ratings for reliability:** 2 high and 4 moderate → Measure passes with MODERATE rating
    - Data used in testing: Data collected from October 2016 through October 2018 regarding care experiences of patients who died while receiving hospice care from July 2016 through June 2018). This included a testing sample of 2,933 hospices and 647,694 surveys.
    - Data element reliability was calculated for the multi-item measures by examining: (1) the internal consistency of the multi-item measures using Cronbach's alpha and (2) the item-total correlation using Pearson's correlation
        - Cronbach's alpha values ranged from 0.61 to 0.84) [lowest for timely care, respect, and emotional/religious support]
        - Pearson correlations values ranged from 0.39 to 0.70
    - The developer did not conduct test-retest reliability assessments for the 2 single-item measures, due to sensitivity of the method to grief and bereavement. However, the developers cited previous literature showing relative high levels of agreement over time for items related to overall quality of hospice care and willingness to recommend the hospice.
    - Measure score reliability was calculated using: (1) intra-class correlations (ICCs) computed from the case mix-adjusted 0-100 top-box scores and (2) estimating reliability via the Spearman-Brown prophecy formula (using the average number of completed surveys per hospice over the eight quarters of data).
        - ICCs ranged from 0.012 0.025
        - Spearman-Brown reliability estimates ranged from 0.71 to 0.85
- **Ratings for validity:** 6 moderate → Measure passes with a MODERATE rating
    - Developers conducted construct validation analyses of the items in the survey (i.e., data element testing) in two ways: (1) by examining the relationship of the individual-level results from the 6 multi-items measures to the individual-level results of the global rating measure and (2) by examining Pearson correlations between the individual-level multi-item measures to assess the magnitude of association between them.
        - Correlations with rating of hospice PRO-PM ranged from 0.42 to 0.61 (all statistically significant)
        - Correlations with willingness to recommend hospice PRO-PM ranged from 0.41 to 0.58 (all statistically significant)
        - Pearson correlations between the multi-item measures ranged from 0.31 to 0.63 (all statistically significant)
    - Developers conducted construct validation analyses of the measure scores in two ways: (1) by examining the relationship of the agency-level results from the 6 multi-items measures to the agency-level results of the global rating measure and (2) by examining Pearson correlations between the agency-level multi-item measures to assess the magnitude of association between them.
        - Correlations with rating of hospice PRO-PM ranged from 0.63 to 0.84 (all statistically significant)
        - Correlations with willingness to recommend hospice PRO-PM ranged from 0.60 to 0.81 (all statistically significant)

        - Pearson correlations between the multi-item measures ranged from 0.42 to 0.84 (all statistically significant)
    - A variety of patient-level exclusions are specified for the measure. However, the developers did not provide information on the frequency/variation of these exclusions.
    - Risk-adjustment
        - Each of the 8 PRO-PM measures are case-mix adjusted using a linear regression model.
        - The same 9 factors are included in the case-mix adjustment approach for each PRO-PM. These include: response percentile (days between death and survey response); decedent age group; payer; primary diagnosis; length of final hospice episode; respondent age group; respondent education; decedent's relationship to respondent; a variable indicating survey/respondent's home language (Spanish vs other).
        - Each of the 8 PRO-PM measures also are adjusted for the mode of survey administration (this is done prior to the case-mix adjustment).
        - Need for case-mix adjustment was assessed via a Kendall's tau, a measure of rank correlation. A tau estimate equal to 1 would indicate that case-mix adjustment has no effect on the hospice-level scores.
            - Tau values ranged from 0.88 to 0.94, indicated a small effect of case-mix adjustment. Developers suggest this adjustment may be important for hospices with unusual case mix.
    - Meaningful differences: depending on the measure, the scores of approximately one-quarter to one third of hospices are statistically significantly different from the national average.
    - Comparability of methods: Developers found that mode of administration does effect responses to the measures and therefore, the measure scores must be adjusted for mode of administration.
    - Non-response and missing data
        - Overall response rate=32.63%
        - Inappropriately missing responses for the various items in the survey ranged from 0.5% to 4.8%. The highest numbers of missing items were for 2 of the 3 emotional/religious support questions.
- Some concerns of the SMP
    - Desire for more clarity regarding how the measure scores are calculated
    - Concerns about selection bias (e.g., for those without caregiver of record)
    - Length of final episode of hospice care is a case-mix variable that is not present at start of care
    - Concern about excluding hospice stays of <48 hours
    - Desire for additional construct validation with measures not derived from Hospice CAHPS

ITEMS TO BE DISCUSSED

- None

Measure# 3533e: Hospital Harm – Severe Hyperglycemia

MEASURE HIGHLIGHTS

- New measure
- **Description:** This ratio electronic clinical quality measure (eCQM) assesses the number of hospital days with a severe hyperglycemic event (a blood glucose result >300 mg/dL, or a day in which a blood glucose value was not documented and it was preceded by two consecutive days where at least one glucose value is >=200 mg/dL) per the total qualifying hospital days among inpatient encounters for patients 18 years and older who have either:

  1. A diagnosis of diabetes mellitus,
  2. Received at least one administration of insulin or an anti-diabetic medication during the hospital admission, or
  3. Had an elevated blood glucose level (>200 mg/dL) during their hospital admission.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Records
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Ratings for reliability:** H-6; M-0; L-0; I-0 → Measure passes with HIGH rating
  - To assess reliability of the measure score, the developers used Adams' beta-binomial method to calculate the signal-to-noise ratio.
    - There were 5,501 eligible encounters (and 19,736 eligible days) across Hospitals 1-6. The signal-to-noise ratio yielded a median reliability score of 0.967 (range: 0.955-0.983).
- **Ratings for validity:** H-4; M-1; L-0; I-1 → Measure passes with HIGH rating
  - Data element validity was assessed by evaluating the accuracy of electronically extracted EHR data elements compared with manually chart abstracted data elements from the same patients, which is considered the "gold standard" for these analyses.

ITEMS TO BE DISCUSSED

- None

Subgroup 2

Measure# 0071: Persistence of Beta-Blocker Treatment After a Heart Attack

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The percentage of patients 18 years of age and older during the measurement year who were hospitalized and discharged from July 1 of the year prior to the measurement year to June 30 of the measurement year with a diagnosis of acute myocardial infarction (AMI) and who received persistent beta-blocker treatment for six months after discharge.
- **Type of measure:** Intermediate Clinical Outcome
- **Data source:** Claims
- **Level of analysis:** Health Plan
- **Not risk-adjusted**
- **Ratings for reliability:** 2 high, 5 moderate → Measure passes with MODERATE rating

- Testing data included HEDIS 2018 plan data (2017 measurement year)
    - 243 commercial plans, median denominator size=65
    - 145 Medicaid plans, median denominator size=81
    - 272 Medicare plans, median denominator size=72
- Score-level reliability testing was conducted using the beta-binomial model described by Adams (2009).
    - Average reliability, commercial: 0.757; 25th percentile=0.521, median=0.672
    - Average reliability, Medicaid: 0.818; 25th percentile=0.389, median=0.621
    - Average reliability, Medicare: 0.730; 25th percentile=0.670, median=0.772
- Data element reliability testing was not conducted. NOTE that such testing is NOT required by NQF for this type of measure.

- **Ratings for validity:** 5 moderate, 1 low, 1 insufficient → measure passes with a MODERATE rating
    - Testing data same as described above
    - Score-level construct validation was conducted by correlating the scores for this measure to those of a measure of statin therapy adherence.
        - Developers hypothesized that a plan that does well on the statin adherence measure for cardiovascular disease would also do well on this measure.
            - Pearson correlation coefficient, commercial: 0.51 (statistically significant)
            - Pearson correlation coefficient, Medicaid: 0.60 (statistically significant)
            - Pearson correlation coefficient, Medicare: 0.42 (statistically significant)
            - Developers interpret these results as supporting their hypothesis and validating this measure.
    - Exclusions
        - This measure includes several exclusions, but except for advanced illness/frailty, the developers did not test the exclusions.
        - Exclusions related to hospice enrollment, I-SNP enrollment, living in a long-term care intuitional setting, and advanced illness and frailty are new since the measure was last evaluated by NQF for endorsement.
            - Exclusions for advanced illness resulted in a loss of 4.6% of patients on average (in 10 plans), and a 2.5% higher performance rate on average across the 10 plans
            - Exclusions for frailty resulted in a loss of 1.1% of patients on average (in 10 plans), and a 0.5% higher performance rate on average across the 10 plans
    - This measure is not risk-adjusted. Developers provide a brief conceptual rationale regarding lack of risk-adjustment.

- To demonstrate ability to identify statistically meaningful differences across health plans:
    - The developers presented distributional statistics by plan type (e.g., average, standard deviations, IQR, etc.)
    - Used an independent sample t-test of the performance difference between two randomly selected plans at the 25th versus 75th percentile. The test statistic is compared against a normal distribution. If the p-value of the test statistic is less than .05, then the two plans' performance is significantly different from each other. P-values for all three plan types were <0.05.
- Missing data
    - The developers describe how their audit process considers missing data. However, they do not present information on the extent of missing data.

- Some concerns of the SMP
    - Unclear why the developers have classified this measure as an intermediate clinical outcome
    - Desire for clarity regarding patients who have <180 days of treatment because they died in that timeframe
    - Would like to have seen more detail regarding the method used to test reliability, as well as more detail regarding reliability in relation to sample size
    - Concern regarding low reliability for many plans (particularly Medicaid plans)
    - Desire for data regarding frequency of exclusions
    - Desire for data characterizing the extent of missing data
    - There is some disagreement about the need for risk-adjustment, given the characterization of the measure as an intermediate clinical outcome (note that risk-adjustment not expected for process measures)
    - Some concern about the utility of the statin measure as a comparator/validator for this measure

ITEMS TO BE DISCUSSED

- None

Subgroup 3

Measure# 3538: All-Cause Emergency Department Utilization Rate for Medicaid Beneficiaries Who May Benefit from Integrated Physical and Behavioral Health Care

MEASURE HIGHLIGHTS

- New measure
- **Description:** The measure focuses on emergency department (ED) utilization for four populations of Medicaid beneficiaries who may benefit from integrated physical and behavioral health care. The rates in this measure are intended to be reported at the state level. This is an inverse measure; lower scores indicate better quality of care.
  The measure is defined as the all-cause ED utilization rate for Medicaid beneficiaries age 18 and older who meet the eligibility criteria for any of the four denominator groups:

1. Beneficiaries with co-occurring physical health and mental health conditions (PH+MH)
2. Beneficiaries with a co-occurring physical health condition and a substance use disorder (PH+SUD)
3. Beneficiaries with a co-occurring mental health condition and a SUD (MH+SUD)
4. Beneficiaries with serious mental illness (SMI)

The measure is calculated over the period of one calendar year as the number of ED visits that do not result in an inpatient admission or observation stay per 1,000 member-months. It is reported as four separate rates, one for each denominator group.

Each of the four denominator groups includes only beneficiaries who were not dually eligible, were enrolled in Medicaid for at least 10 months of the measurement year, and had a diagnosis within the measurement year or year prior (depending upon the condition) that placed them into one or more of the denominator groups.

- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Population: Regional and State
- **Risk Adjusted:** Statistical risk model, observed vs. expected weighting
- **Ratings for reliability:** 5 high and 1 moderate → Measure passes with HIGH rating
  - Measure score reliability performed with signal-to-noise analysis
- **Ratings for validity:** H-2; M-4; L-0; I-0→ Measure passes with a MODERATE Rating
  - Measure score level validity testing performed

Table 4. Spearman rank correlation between this measure and five Core Set measures

Core Set Measure*	PH + MH	PH + SUD	MH + SUD	SMI
FUH	0.25 (-0.38, 0.72)	0.17 (-0.44, 0.68)	0.43 (-0.19, 0.81)	0.31 (-0.32, 0.75)
IET	0.48 (-0.22, 0.85)	0.64 (0.01, 0.9)	0.36 (-0.35, 0.81)	0.6 (-0.05, 0.89)
SAA	0.83 (0.42, 0.96)	0.66 (0.05, 0.91)	0.81 (0.36, 0.95)	0.43 (-0.27, 0.83)
MPM	-0.04 (-0.58, 0.52)	-0.03 (-0.57, 0.53)	-0.27 (-0.72, 0.33)	-0.13 (-0.63, 0.46)
AMM	0.27 (-0.39, 0.75)	-0.07 (-0.64, 0.55)	0.42 (-0.24, 0.81)	-0.19 (-0.71, 0.46)"

* Acronym definitions:
Follow-Up After Hospitalization for Mental Illness: Age 21 and Older (**FUH**), 7-Day Rate
Initiation and Engagement of Alcohol and Other Drug (AOD) Abuse or Dependence Treatment (**IET**), Initiation of AOD Treatment Rate
Adherence to Antipsychotic Medications for Individuals with Schizophrenia (**SAA**)
Annual Monitoring for Patients on Persistent Medications (**MPM**)
Antidepressant Medication Management (**AMM**), Acute Phase Treatment Rate

- **Exclusions:** Medicare duals (so how were there folks over 65?), also < 10 months of Medicaid in the year. Not tested.
- **Risk adjustment approach:**
  - Steps summarized methods here**:** models use to explore factors that influence the number of ED visits: negative binomial chosen, over Poisson and ZINB, without strong explanation. Backward selection was used to remove ns factors. Unified model for all 4 groups developed together, even as each group may respond differently. Table 5 p.17 has raw coefficients for the 57 variables. Used Andersen's Behavioral Model (enabling factors, etc.), but income/education not included because Medicaid is used as a proxy for low income.
    - Largest effect was fibromyalgia and migraines (IRR>1.7 and 1.57 , respectively)

- **Statistically meaningful differences:**
  - Across 17 states:
  - PH+MH: range—175/1,000 to 265/1,000; 13 of 17 states were statistically different from the overall average. 7 more than average "suggesting room for improvement".
  - PH+SUD: range—234 to 378; as above but 6 above average
  - MH+SUD: range—207 to 323; 14 states were statistically different than average and 7 were more than average
  - SMI: range—229 to 362; 12 states were sig different than average and 5 were higher….
- **Missing data:** Typically less than 1% of the data was reported missing

ITEMS TO BE DISCUSSED

- None

Measure# 0684: Percent of Residents with a Urinary Tract Infection (Long Stay)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure reports the percentage of long-stay residents who have a urinary tract infection in the 30 days prior to the target assessment. This measure is based on data from the Minimum Data Set (MDS) 3.0 OBRA, PPS, and/or discharge assessments during the selected quarter. Long-stay nursing facility residents are identified as those who have had 101 or more cumulative days of nursing home care.
- **Type of measure:** Outcome
- **Data source:** Assessment Data
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Testing data**
  - The data set used for testing was the Nursing Home Minimum Data Set (MDS) 3.0 v1.15.0
  - Two studies were used for the analysis:
    - The RAND Development and Validation of MDS 3.0 study sample included a representative sample of for-profit and not-for-profit facilities, and hospital-based and freestanding facilities, which were recruited for the study. The sample included 71 community nursing facilities in 8 states and 19 Veterans Affairs (VA) nursing homes (Saliba & Buchanan, 2008).
      - Included 3,822 residents from community nursing homes and 764 residents from VHA nursing homes
    - RTI facility-level analyses of MDS 3.0 data sample included all facilities with sufficient sample size (n ≥ 20 residents) to publicly report this measure in Quarter 3, 2018 (k = 14,520), unless otherwise noted (RTI International, 2019)
      - Included 1,096, 778 long-stay residents
- **Ratings for reliability:** 5 moderate and 1 low → Measure passes with MODERATE rating
  - To test critical data element reliability, the developers examined agreement in coding of the relevant MDS items between 'gold standard' (research) nurses and facility nurses.

        - The Kappa for gold-standard to facility-nurse agreement on the MDS 3.0 and MDS 2.0 item was 0.70. Kappa is a statistical measure of inter-rater agreement for qualitative data, ranging from 0.0 to 1.0. A rating of 0.70 is considered "substantial agreement."
    - To test reliability of the measure score, the developers performed:
        - A signal-to-noise analysis
            - The signal-to-noise ratio for this measure was 0.191 ($p < 0.001$) indicating that 19.1% of the variance in scores for this measure in Quarter 3, 2018 was explained by inter-facility characteristics (including the underlying quality of care in each facility)
        - A split-half reliability analysis
            - The split-half correlation for this measure was positive, but the relationship was moderate ($r = 0.42$, $\rho = 0.37$, $p < .001$), and the ICC was 0.42 ($p < .001$)
- **Ratings for validity:** 1 high, 3 moderate, 1 low, 1 insufficient → Measure passes with a MODERATE rating
    - To assess validity of the measure score, the developers examined whether a facility's percentile rank on this measure was correlated with its percentile rank on the related quality measures NQF #0686 (Percent of Residents Who Have/Had a Catheter Inserted and Left in Their Bladder (Long Stay)) and NQF #0685 Percent of Low Risk Residents Who Lose Control of their Bowel or Bladder (Long Stay)).
    - As additional support for measure validity, the developers also analyzed:
        - Variation by state
        - Seasonal variation
        - Stability over time (change in percentile ranking in consecutive quarters and average change in performance across years)
        - Face validity
            - a Technical Expert Panel (TEP) provided feedback on the face validity of NQF #0684. TEP members discussed the current measure specifications, potential risk adjustment factors, and the effectiveness of the measure in capturing quality of care to determine the face validity of the measure as it is currently specified.
- Methods Panel Concerns:
    - One reviewer found the specifications to be unclear with regard to the measurement timeframe
    - Panel members suggested that testing results showed strong reliability at the data element level but only moderate reliability at the measure score level.
    - Some concerns were raised about the adequacy of the developers' rationale for not risk adjusting the measure and the lack of any testing of risk adjustment models.

ITEMS TO BE DISCUSSED

- None

Subgroup 4

Measure# 2979: Standardized Transfusion Ratio for Dialysis Facilities

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The risk adjusted facility level transfusion ratio "STrR" is specified for all adult dialysis patients. It is a ratio of the number of eligible red blood cell transfusion events observed in patients dialyzing at a facility, to the number of eligible transfusion events that would be expected under a national norm, after accounting for the patient characteristics within each facility. Eligible transfusions are those that do not have any claims pertaining to the comorbidities identified for exclusion, in the one year look back period prior to each observation window. This measure is calculated as a ratio, but can also be expressed as a rate.
- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- Adjusted/Not risk-adjusted: Adjusted (Statistical Risk Model)
- **Ratings for reliability:** 2 high, 3 moderate, 1 low → Measure passes with MODERATE rating
    - Score-level reliability: Tested the reliability using bootstrapping to evaluate inter-unit reliability (IUR). The results demonstrated a moderate level of reliability.
- **Ratings for validity:** 4 high, 2 insufficient → measure passes with HIGH rating
    - In general, SMP reviewers satisfied with exclusions and risk adjustment approach and methodology.
    - Score-level validation: Face validity assessed using TEP. Empirical testing used Poisson regression model demonstrated association with hospitalization, mortality, and percent of patients with low hemoglobin levels.

ITEMS TO BE DISCUSSED

- None

Measure# 3543: Patient-Centered Contraceptive Counseling (PCCC) measure

MEASURE HIGHLIGHTS

- New measure
- **Description:** Brief description: Instrument with 4 questions.

Ask women age 15-45 whether their contraceptive counseling was patient-centered based on likert scales. A "top box" approach to scoring is used which means that only a perfect "5-excellent" on all items is counted as a numerator event. The items are:

1. Respecting me as a person
2. Letting me say what matters to me about my birth control
3. Taking my preferences about my birth control seriously
4. Giving me enough information to make the best decision about my birth control method

- **Type of measure:** Outcome: PRO-PM
- **Data source:** Patient survey responses, San Francisco and 9 other regions (several in OR), data from 2009 to 2017.

    - 44% MDs, NPs, certified nurse midwife, or PA, 29.4% non-licensed medial assistants, 17.7% two person teams with and NP.
    - 2,200-2,400 providers sampled, for reliability and validity, unclear why different numbers
    - 3,000-3500 facilities sampled, for reliability and validity, unclear why different nubers
    - 341 patient observations and 38 providers (how selected?) used for element level testing.
- **Level of analysis:** Clinician : Individual; Facility
- **Not risk-adjusted**
- **Ratings for reliability:** 5 high and one moderate → Measure passes with HIGH rating
    - Cronbach's alpha for element level 0.94 (items ranging from 0.8-0.89); Facility level 0.93 0.78-0.87)
        - SNR for score level (Spearman-Brown ICC reliability (0.85 for facility, 0.84 for provider)
            - ICCs or 0.1-.15 yielded S-B reliabilities of 0.81 (0.63, 092) or 0.74 (0.59, 0.86) at the provider and facility levels direction; for panels sizes of 25. See table 7.
- **Ratings for validity:** H-5; M-1; L-0; I-0 → Measure passes with HIGH
    - Face validity with modified Delphi methods including providers, administrators, and patients.
    - Data elements (i.e., each of the four questions) were compared again Four Habits Coding Scheme (4HCS) of audio tapes of the same visit. The 4HCS is a validate way to assess provider communication. Comparison made with linear mixed model and random effects (clustering by provider).
    - Performance score: comparing the PCCC measure to two separate measure of patient satisfaction about birth control choice. Make point that satisfaction scores are less specific than their measure, but still another way to assess patient centered contraception care.
    - "high PCCC scores are positively associated with method choice satisfaction (r=0.82) and satisfaction with provider help with birth control choice (r=0.88) both p<0.001. Likewise, at the facility-level, aggregated PCCC scores are associated with r=0.76 with method choice satisfaction and r=0.82 for satisfaction with provider help (n=15 facilities; both p<0.001)."
    - Meaningful differences**:** Providers and facilities, respectively: medians 86 and 83%, 25^{th}/75^{th} percentiles: 75/90% and 70/88%. The also note that available states on the CG-CAHPS (clinician and group, Consumer Assessment of Healthcare Providers and Systems communication composite score) had 88% and 84/91% respectively.
    - Missing data: The 50% of those at each site who received counseling were simpal age and race to those not so counseled
        - Missing data was assess at 1.7% with some entities having missing data as high as 10%. Obliquely explained imputation analyses was said to demonstrate these empty cells did not alter inferences.

ITEMS TO BE DISCUSSED

- None

Appendix B: Additional Information Submitted by Developers for Consideration

Measure Number: 0018

Measure Title: Controlling High Blood Pressure

Measure Developer/Steward: National Committee for Quality Assurance

---

Reliability

- **Issue 1:** The age range in the response for the *S.6 Denominator Statement* question contains a typo.

    - **Developer Response 1:** The correct response should read "Patients **18-85** years of age who had at least two visits on different dates of service with a diagnosis of hypertension during the measurement year or the year prior to the measurement year." This should clear up the confusion created by the discrepancy between the denominator statement, measure description, and subsequent age range statements about exclusions.

- **Issue 2:** One reviewer noted the following:
  "The patient is not compliant if the blood pressure is =140/90 mm Hg.

  The patient is not compliant if the BP reading is =140/90 mm Hg or is missing…
  I assume there is a typo – i.e., ≥?"

    - **Developer Response 2:** Correct. This appears to be a typo that occurred while transferring information into the online Intent to Submit form. The statements should read:
      "The patient is not compliant if the blood pressure is ≥140/90 mm Hg."
      "The patient is not compliant if the BP reading is ≥140/90 mm Hg or is missing…"

- **Issue 3:** One reviewer noted the following:
  "There are exclusion criteria described under the numerator details that are not included in the denominator exclusions. For example:
    - Do not include BP readings:
    - Taken during an acute inpatient stay or an ED visit.
    - Taken on the same day as a diagnostic test or diagnostic or therapeutic procedure that requires a change in diet or change in medication on or one day before the day of the test or procedure, with the exception of fasting blood tests.
    - Reported by or taken by the patient."

It is preferred that all exclusions are detailed within the S.10. Denominator Exclusions section, e.g., Two telehealth visits, Type of visits (Diagnostic tests), self-report."

- **Developer Response 3:** We provide this guidance to health plans because there are often multiple blood pressure readings for patients who qualify for the denominator of this measure. Some readings are taken during situations that may elevate the patient's blood pressure or result in inaccurate readings. Rather than remove the patient altogether by excluding them, we provide guidance to the health plans for what type of clinical encounters and results should not be used for the numerator.

- **Issue 4:** One reviewer noted, "it may be useful to explain why and how a sample of 411 medical records per plan was selected for testing. If all plans assessed exactly 411 medical records, why was the median reported and not simply the number of records? It would be useful to describe how many cases were assessed including claims and medical records per product line, e.g., count, mean, median, min, max."

  - **Developer Response 4:** The data used for the reliability and validity analyses for this submission came from health plans who reported the HEDIS measure to NCQA for measurement year 2018. As described in section 1.6, most health plans use a combination of data from administrative claims and a random sample of 411 medical records they review to report performance rates. However, there are some health plans that report on the full population that qualifies for the denominator through administrative claims and there are some health plans that have fewer than 411 members who qualify for the denominator. This means that there is a range of denominator sizes reported to NCQA every year for this measure, but the median is generally 411.

    Additionally, NCQA maintains detailed guidelines on the calculations and sampling that are used by health plans to report the measure, how to draw the sample of 411, guidance for oversampling when necessary, and how to handle denominators that are less than 411.

- **Issue 5:** One reviewer noted, "The methods used for score reliability testing cannot be fully assessed - Beta-binomial model (Adams, 2009), but no details were provided on the actual methods and formulas used. More details on the statistical method and specific formulas used to calculate the proportion of variability in measured performance that can be explained by real differences in performance would be helpful to better understand what was done exactly."

  - **Developer Response 5:** Our general approach for assessing reliability is provided in section 2a2.2. We used signal to noise reliability to estimate the reliability of pass/fail measures. We calculated within and between plan variance to ensure that the variation between plan rates is significantly large enough to override variance we observed within

plan rates, using 0.70 as the threshold to indicate a signal; in other words, that the following is true:

- Variance Between Plans / (Variance Between Plans + Variance Within Plans) > 0.70

Variance between plans is a function of the distribution of plan-level rates. Variance within a plan is a function of performance rate and sample size ((p(1-p)/n)).

Validity

- **Issue 1:** One reviewer stated the following:
  "In general, the method for determining exclusions was principled and systematic. S.10. The following is listed as an exclusion "patients who had a nonacute inpatient admission during the measurement year" but no rationale is given."

  - **Developer Response 1:** We exclude patients who had a nonacute inpatient admission because we recognize that this population is likely experiencing a more complicated clinical situation during the measurement year.

- **Issue 2:** One reviewer stated the following:
  "I would appreciate learning the rationale for
  - including a patient with one office visit followed by one telephone visit.
  - excluding patients with renal disease/nephrectomy manifesting by Dec 31 of the assessment year and the rationale,
  - excluding pregnant women as blood pressure management is clinically important for many of these patients (perhaps there is a different measure for such patients or determination of pregnancy is problematic.)"

  - **Developer Response 2:** We will respond to this set of comments in the order they were provided:
    - Following an analysis focused on the use of telehealth in healthcare delivery, which included a review of literature and guidelines as well as vetting with expert advisory panels, we determined that there is adequate evidence to support the use of telehealth services in the routine management of patients with hypertension. Therefore, one of the two qualifying encounters for the denominator is permitted to occur through telehealth. Please note that this is not a requirement. We are simply allowing a telehealth encounter, if it occurs, to qualify for one of the two encounters required for denominator identification.
    - Patients with renal disease/nephrectomy are excluded because their specific clinical situation typically requires more individualized care and a blood pressure goal of <140/90 mm Hg may not be appropriate.

        - Like the abovementioned rationale, pregnant women may require a more individualized treatment plan than the general population of people with hypertension. Therefore, we exclude them from this accountability measure.

- **Issue 3:** Several reviewers noted that there is likely an overlap of the patients that qualify for the denominator populations for the Controlling High Blood Pressure measure and the Comprehensive Diabetes Care – Blood Pressure Control measure used in the empirical validity analysis. This overlap combined with the exact same numerator focus of blood pressure control (<140/90 mm Hg) may be causing "somewhat of a circular test of the same data element measure for basically two overlapping samples." One review specifically noted concern that "This assessment is not very strong since the diabetes population is essentially a subset of the overall population. They are essentially assessing exactly the same measure in a subset of the population."

    - **Developer Response 3:** The samples used for the empirical validity analysis come from data that health plans reported for the two measures separately in measurement year 2018; in other words, the same samples were not used for reporting the denominator populations. However, we do recognize there is an inherent overlap between the populations that qualify for denominators focused on hypertension and diabetes. Therefore, NCQA conducted an additional empirical validity analysis, using the Pearson's correlation test, comparing the Controlling High Blood Pressure measure to the following measures:

        - Comprehensive Diabetes Care: HbA1c Control (< 8%): The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) whose most recent HbA1c level is < 8.0% during the measurement year.
        - Comprehensive Diabetes Care: HbA1c Poor Control (> 9%): The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) whose most recent HbA1c level is > 9.0% during the measurement year.

      These measures were chosen for construct validity testing because they are similarly focused on the management of a chronic condition but aimed at different biological markers. We hypothesized that a plan that does well on one measure focused on the management of blood pressure for patients with hypertension will likely do well on other measures focused on the management of other chronic conditions, such as blood glucose for patients with diabetes. **Note:** The HbA1c Poor Control measure is a "lower is better quality" measure. This means that plans that are performing well will have low rates on this measure.

      The results were the following:

Table 1. Correlations between CBP and CDC HbA1c measures in Commercial Health Plans, 2018.

	Pearson Correlation Coefficients	
	CDC – HbA1c Control	CDC – HbA1c Poor Control
CBP	0.810	-0.824

Note: All correlations are significant at p<0.0001

Table 2. Correlations between CBP and CDC HbA1c measures in Medicare Health Plans, 2018.

	Pearson Correlation Coefficients	
	CDC – HbA1c Control	CDC – HbA1c Poor Control
CBP	0.519	-0.577

Note: All correlations are significant at p<0.0001

Table 3. Correlations between CBP and CDC HbA1c measures in Medicaid Health Plans, 2018.

	Pearson Correlation Coefficients	
	CDC – HbA1c Control	CDC – HbA1c Poor Control
CBP	0.795	-0.820

Note: All correlations are significant at p<0.0001

Across all product lines, these correlations are moderate to very strong and statistically significant.

Measure Number: 0425

Measure Title: Functional Status Change for Patients with Lumbar Impairments

Measure Developer/Steward: Focus On Therapeutic Outcomes, Inc (FOTO)

Reliability

- **Issue 1:** Panel member 4 asked whether this measure reaches the levels of reliability the committee expects for assessing the performance of clinicians with 10-19 and 20-29 cases. It was also noted that proportion of variance was not reported.

    - **Developer Response 1:** We would like to clarify that the reliability reported was for n ≥ (noted as 'X'+) and not 10-19 or 20-29 as interpreted. For example, at the clinician level, the average reliability of 0.71 is for all clinicians that had a least 10 patients per year (n=12,025). As noted, 58% of those clinicians had a reliability above 0.7, with a max reliability of 0.98. Obviously, setting a higher bar for minimum patients required would improve the average reliability of the clinician level. However, this would also cause many clinicians to be excluded as described in the 'N providers' column of TABLE 2a2.3iV, decreasing the ability of NQF measure 0425 to represent performance of the

'real-world'. We think that maintaining the threshold of 0.7 for average reliability of the provider level is an appropriate balance between reliability and provider inclusion. Proportion of variance explained by the provider level from the HLM is shown in column 'Variance explained (%) by the provider level' of TABLE 2a2.3iV. If this was not what the panel member wanted to see, please provide more informative instructions.

Validity

- **Issue 1: Panel member 1 asked the following question regarding structural validity: "**In section 2b1.3ii, I would have liked to have seen the items that were removed and their fit statistics/residual correlations as well as where they fit on the latent construct (person/item map)."

    - **Developer Response 1:** Here is the information provided in the 2006 publication (reference #23):

"In CFA, a two-factor model fit better than a one-factor model, but the correlation between the two factors was high (0.76) suggesting one dominant factor. Fit statistics from the one- to two-factor models were CFI=0.73 and 0.80, TLI=0.97 and 0.98, and RMSEA=0.14 and 0.12, for one- and two-factor models, respectively. Factor loadings for the one-factor solution were all >0.67. These statistics represent mixed results regarding fit for the model. Although the percentage of item variance accounted for was high, factor loadings were adequate, and the magnitude of coefficients was strong, only the TLI index was acceptable for the one-factor solution. Assessment of factor loadings and residual correlations suggested the item sleeping did not load well on the dominant factor and was deleted. Plus, the number of absolute residuals greater than 0.10 was higher than desired suggesting possible local dependence between items related to item pairs (i.e., BPFS items assessing similar tasks like driving for 1 hour and sitting for 1 hour). We decided to test model fit for a 26-item pool without the BPFS item sitting for 1 hour. The number of negative residual correlations was reduced, and the 26-item pool demonstrated a slight improvement in fit statistics (CFI=0.83, TLI=0.98, and RMSEA=0.11 for the one-factor solution). Review of the absolute negative residuals greater than 0.10 for the 26-item pool revealed possible local dependence between apparently unrelated items (i.e., BPFS items assessing walking a mile and putting on shoes and socks). We decided to test model fit for a 25-item pool without the BPFS item walking a mile. The number of negative residual correlations was eliminated, and the 25-item pool demonstrated a slight improvement in the fit statistics (CFI=0.87, TLI=0.98, and RMSEA=0.09 for the one-factor solution). We believe the 25-item pool represented a unidimensional pool with strong local independence."

A person/item map of the item pool was not provided as part of the 2006 paper but could be recreated if needed.

- **Issue 2:** Panel member 1 had concerns over the percent of explained variance of the risk-adjusted model (37%) being low.

    - **Developer Response 2:** From our experience with predictive models of functional status PROMs, and existing literature, we interpret that explaining 30-40% of variance in outcomes using only baseline patient characteristics that are outside of the provider's influence, excluding treatment processes, is considered high. We agree that additional variables should always be considered, as done for this maintenance submission compared to the original submission, where we updated and improved the risk-adjustment model. We plan on continuing to carefully reassess the model, also considering the risk of over-adjusting for factors that could potentially be influences.

- **Issue 3:** Panel member 1 wanted to have more scale level psychometric information.

    - **Developer Response 3:** This information is included in the 2006 development paper, including reasons for selecting the RSM for item calibration (although not described in the paper, the partial credit model did not provide better fit compared to the RSM), Item characteristics of the 25-item bank (Table 2), scale level fit, reliability and effective range (Table 2), and number of CAT items used with their corresponding standard errors by level of ability (Figure 1). We felt these descriptions were beyond the requested information for this submission therefor they were not included and only referenced.

- **Issue 4:** Panel member 2 noted that we did not list an exclusion of patients without 2 assessments (those who didn't complete rehab or didn't complete both admission and completion assessment).

    - **Developer Response 4:** This is correct although this is an obvious exclusion criterion following the numerator definition. We suggest adding this as an additional exclusion criterion for clarity.

- **Issue 5:** Panel member 2 asked about the possibility that improvement of provider's performance over time (those that provided data for all three years) could be a result of confounding introduced by PROMs completed by proxies.

    - **Developer Response 5:** As noted in the measure information form (S.15), proxy data were very rare (0.03%), therefore it is unlikely that it could have impacted provider's performance results. As described, this was also the reason for not assessing proxy data separately in our analyses.

- **Issue 6: Panel member 2** asked if patients excluded resulted in bias in measurement, specifically in relation to social risk factors or use of proxy to complete the PROM.

    - **Developer Response 6:** As mentioned above, proxy data were very rare and did not justify or allow any testing to be conducted as to potential bias introduced by PROMs completed by proxy, or use of proxy data within the risk-adjustment model. As to the possibility of clinicians responding to the low-back PROM on behalf of the patient, although we cannot ascertain that this never happens, this is strictly unadvised by FOTO as part of the FOTO standards of PROM administration. A systematic bias introduced by such unadvised practice would have also impacted our missing data results. These included testing for potential bias related to missing outcomes data using three methods. 1) comparing patients with or without complete outcomes, 2) assessing correlations between clinician and clinic residuals and their completion rates, and 3), assessing average residuals at the clinician or clinic levels by completion rate categories with or without an adjustment using inverse probability weighting. These methods incorporated payer data that could to some extent serve as a proxy for socio-demographic information, although we agree that payer does not fully represent social-risk factors.
      No evidence of a systematic selection bias was observed.
      Other social risk-factors available to us were education levels for a sub-sample of patients, with only negligible impact on the predictive power of the risk-adjusted model. We did not include a patient comparison between those with complete or incomplete outcomes by educational level for the sub-sample of patients that had educational level data, since education level is not a standard data element that is being collected (TABLE 2b6.2i). This information is added in the following table. Overall, distribution of educational levels was very similar between those with complete or incomplete outcomes data, although higher educational levels (Bachelor's or above) were slightly more frequent in the complete outcomes group. Overall, although some selection bias might exist between patients with or without complete outcomes data, we interpret our results as no evidence to support a bias that would pose a threat to NQF 0425 measure's validity. We agree with the comment that additional testing of collapsed educational levels as a risk-factor is a logical next step for testing social risk-factors that may or may not be advised for inclusion in our risk-adjusted model, as we noted in the submission (2b3.4b).

Level of education	**Complete outcomes* N=41,889**	**Incomplete outcomes* N= 19,024**	**Total* N=60,913**
Less than high-school degree	5.7	6.8	6.0
High-school degree or equivalent	22.4	23.8	22.8
Trade/technical/vocational training	5.9	6.5	6.1
Some college but no degree	16.5	17.0	16.7

Level of education	Complete outcomes* N=41,889	Incomplete outcomes* N= 19,024	Total* N=60,913
Associate degree	8.3	8.5	8.4
Bachelor's degree	20.0	18.2	19.4
Master's degree	10.9	9.6	10.5
Other advanced degree beyond a Master's	4.2	3.8	4.1
Prefer not to answer	6.1	5.9	6.1
*Values are in percent			

- **Issue 7: Panel member 4** suggested to assess a possible association between the risk-adjusted construct of acuity (time from condition onset to admission) and SES factors as access to providers. Also, a concern was raised regarding the method used to assess education as a risk-factor.

    - **Developer Response 7:** We agree that access to rehabilitation may be an important factor to test and consider for risk-adjustment. Unfortunately, this information is not currently included within our data. We plan to test feasibility of SES data collection in the future, to enable a more comprehensive assessment of social risk-factors.
      We would like to clarify that education level was tested as one construct with multiple categories, and not as a separate construct for each education level as may have been interpreted by the panel member. With a stepwise approach (entry=0.005; removal=0.01), the educational construct would have been retained in the model. However, all categories combined had a sum of squared semi-partial of 0.02%, suggesting minimal added value to the overall predictive strength of the risk-adjusted model. As we noted in the interpretation section, these findings were considered preliminary and we plan on continuing to test the education construct using different groupings. As reported, and can also be interpreted using the unstandardized beta coefficients shown in the table below, there was no consistent pattern of higher outcomes for higher educational levels.

Variables	Frequency	B	T	Sig	Squared Semi-partial correlation
(1)Less than high school degree (Reference group)	6%				
(2)High school degree or equivalent	22%	-0.4	-1.38	0.167	0.00%
(3)Trade/technical/vocational training	6%	-0.1	-0.37	0.710	0.00%

Variables	Frequency	B	T	Sig	Squared Semi-partial correlation
(4)Some college but no degree	16%	0.0	-0.04	0.968	0.00%
(5)Associate's degree	8%	-0.1	-0.22	0.826	0.00%
(6)Bachelor's degree	20%	0.7	2.23	0.025	0.01%
(7)Master's degree	11%	0.5	1.55	0.120	0.00%
(8)Other advanced degree beyond a Master's	4%	0.4	0.95	0.343	0.00%
(9)Prefer not to answer	6%	0.8	1.97	0.049	0.01%

- **Issue 8: Panel member 5** suggested adding an upper age limit as an exclusion criterion.
    - **Developer Response 8:** Currently, FOTO does not limit the completion of PROMs for the elderly population. Since HIPAA regulations have been implemented, all patients above the age of 89 are grouped and de-identified for their exact age. Practically, this allows assessing a continuous age variable up to the age of 89, which could be considered as the upper age limit for NQF measure 0425.

Other General Comments

Specifications:

Issue 1: From panel member 1: *"in the sampling section S.15 there are instructions for patients less than and 8 years old and for those over 8. I found this confusing since the measure description in DE3 indicated that the measure is to be used on those 14 and over."*

- **Developer Response 1:** We understand how this information may have been confusing. These are the general age related instructions for use of proxy (under age 8 or above 8 but uncomfortable responding). However, any age below 14 will not be included in measure 0425. We could add a note to S.15 to clarify this.

Issue 2: From panel member 2: *"The denominator statement does not define "who have initiated an episode of care"? How do they identify this population? The ICD codes for low back impairment are listed. Also "and who completed the low back FS PROM". Don't they mean at least twice? Does it have to be at initiation of the episode (admission?) and at discharge? How are those two specific patient reported outcome surveys identified?"*

- **Developer Response 2: We will be happy to provide clarification to the Denominator Statement and the Denominator Details (i.e., the location of the ICD-10 codes).**
  Denominator Statement: The target population is all patients 14 years and older with a Low Back impairment who completed an episode of care and completed the Low Back FS PROM at the time of Initial Evaluation and Discharge.
  Denominator Details: The target population is all patients 14 years and older with a low back impairment and/or diagnosis pertaining to a functional deficit affecting the low back: (list of ICD-10 codes)

Measure Number: 0684

Measure Title: Percent of Residents with a Urinary Tract Infection (Long Stay)

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

Reliability

- **Issue 1:** Ambiguity in definition of long-stay denominator for NHC measures.
  - **Developer Response 1:** Chapter 1 Section 1 of the MDS 3.0 QM User's Manual https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-instruments/NursingHomeQualityInits/NHQIQualityMeasures.html describes the detailed methodology used to select the long stay samples. A long stay is defined as an episode with greater than or equal to 101 cumulative days in facility as of the end of the span of time that defines the quality measure reporting period. An episode ends at the earliest of the following:
    - A discharge assessment with return not anticipated (A0310F = [10]), or
    - A discharge assessment with return anticipated (A0310F = [11]) but the resident did not return within 30 days of discharge, or
    - A death in facility tracking record (A0310F = [12]), or
    - The end of the span of time that defines the quality measure reporting period.
- **Issue 2:** The measure includes all nursing home stays, some of which would be private pay or private insurance.
  - **Developer Response 2: This is correct, and is by design.** The assessment requirements for the MDS Resident Assessment Instrument (RAI) are applicable to all residents in Medicare and/or Medicaid certified long-term care facilities. The requirements are applicable regardless of payment source or payer source. Further, these requirements extend to all assessment-based NHC measures; consequently, NHC captures performance data for all stays in Medicare and/or Medicaid certified facilities meeting the case minimum, regardless of whether the stay is covered by private insurance.
- **Issue 3:** The 2008 RAND testing is 11 years old. How similar or different is the MDS 3.0 v1.15.0 compared to MDS 3.0 in the RAND study?

- **Developer Response 3:** The MDS 3.0 item set has remained stable since RAND created the recommended MDS 3.0 form in 2008, with the exception of specific changes in item specification and additions of some new items. In particular, the UTI item has the same look-back period and the same item wording in MDS 3.0 v1.15.0 and the 2008 recommended form.

- **Issue 4:** In the RAND study the short and long term cases were mixed. However, this UTI measure is for long term cases.

  - **Developer Response 4:** As the panelist notes, when selecting residents for the national test, the RAND team aimed to capture a representative sample of short- and long-stay residents together. More broadly, the team's development process intended to ensure the clinical appropriateness of data elements for both short- and long-stay residents. For example, during the MDS revision, the RAND research team asked a national panel of NH experts to rate the utility/importance of core MDS constructs for (a) the clinical care of a person requiring basic nursing facility services, (b) the clinical care of a person requiring skilled nursing or rehabilitation after an acute illness, (c) costs or resource use, and (d) understanding facility quality. In final voting on the core constructs in the MDS, panel ratings of the overall mean clinical importance of the 52 constructs did not differ significantly for long-stay residents vs. short-stay (primarily post-acute care) residents.

- **Issue 5:** Although the referenced Saliba article mentions high reliability, panel member could not locate the exact value.

  - **Developer Response 5:** The quoted numbers are from Table 13.2 and Table 13.3 of the Saliba & Buchanan, 2008 RAND report. http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf.

Validity

- **Issue 1:** Meaningful differences are harder to establish given the narrow distribution of provider scores, but in broad sense meaningful differences are clearer between ends of distribution.

  - **Developer Response 1:** We agree that meaningful differences exist at the ends of the distribution for providers' measure performance. Measure scores are presented on the Nursing Home Compare website such that consumers are able to compare a provider's individual performance with both the state and national average provider scores. The testing form specifies a confidence interval analysis used to identify the proportion of facilities with a score statistically significantly different from the national average. This analysis showed that 39.5% of facilities have a score statistically significantly different

from the national average, including 513 facilities who have a score statistically significantly worse than the national mean. So while a portion of the distribution may be narrow, consumers are still presented information in a way that contextualizes the provider's performance and allows them to meaningfully distinguish a sizeable number of above average and below average performers.

- **Issue 2:** The analyses referenced in the testing form are nearly 15 years old, and the results from Table 1 are more than 5 years old. While the trends are interesting, they only cover 2 years from over 5 years ago.
    - **Developer Response 2:** For clarification purposes, the data from Tables 1a and 1b from the testing form comes from MDS data from Q3 of 2018. Similarly, if the commenter was referring to Figure 1, we would like to further clarify that the analysis covers quarterly trends for over 7 years, with data up to Q3 of 2018 included.
- **Issue 3:** The testing form establishes correlation between the outcome and certain risk factors, including demographic and functional status indicators, suggesting they could be included in a risk-adjustment model to improve the measure performance.
    - **Developer Response 3:** As noted in the testing form, the TEP recommended examining certain functional status indicators, which were hypothesized to be correlated with the outcome measure, including hospice care, bed mobility, transfer, walk in room, walk in corridor, and toilet use. The testing form details an analysis performed to examine the strength of the relationship between each of these indicators and the outcome measure. For each functional indicator, the results suggested very weak correlation with the outcome measure, including non-monotonic relationships between each indicator and measure score deciles. With respect to the social risk factors, CMS has long-established guidance to ensure disparities in care associated with certain social risk factors are transparent to the public. Traditional risk-adjustment models can indeed have the unintended consequence of encouraging such disparities across social risk factors. To combat this issue, CMS favors an approach of reporting measures stratified by the relevant social risk factors, as this allows consumers to observe performance within each relevant group and avoid masking disparities in care. However, stratification of results by group exacerbates the effects of the case minimum for public reporting. Stratifying results by gender, for example, would cut the number of facilities eligible for public reporting by 50%, thus consumers would receive substantially less visibility into provider performance with this stratification.
- **Issue 4:** The testing form includes an article indicating catheter use as a variable influencing UTI onset, which suggests it should be used in risk-adjustment.

- **Developer Response 4:** While the testing form indicates moderate correlation between the UTI and Catheter Insertion (NQF #0686) measures, this result was primarily intended to illustrate the validity of the UTI measure. Including Catheter Insertion in a risk-adjustment model for the UTI has the potential to be problematic due to the fact that providers have substantial control over this process. Risk-adjusting for processes controlled by facilities may allow facilities to game measures.

- **Issue 5:** The stability analysis and change in performance across years would not necessarily tell whether a measure is valid.

    - **Developer Response 5:** We agree that the main purpose of the stability analysis and change in performance across years is to establish the measure's reliability. However, these tests also suggest there is a clear, systematic aspect of quality of care being delivered. Given this and the fact that the underlying data elements accurately capture UTI in residents, the measure accurately reflects systematic features of facilities' performance on UTI rates (rather than conflating other clinical factors).

- **Issue 6:** Test results show the national average score is virtually 0, with 0 being the value for percentiles 10-90. Why should the measure continue when there is no differentiation between providers?

    - **Developer Response 6:** For clarification purposes, we would like to point out that the national average performance in Q3 of 2018 is 2.8% and the median provider score is 1.9%. While 32.3% of facilities had a perfect score of zero on the measure, more than two thirds of facilities had positive scores. Further, the measure identifies systematically poor performers; for instance, as mentioned in Developer Response 1, 513 facilities were identified that performed statistically significantly worse than the national mean. As discussed in the response to Issue 1, the purpose of this measure is to help beneficiaries understand the infection risks associated with different facilities so that they can make an informed decision. So while some providers may have a perfect or near-perfect score, the ability to distinguish poor-performing facilities on the UTI measure continues to have substantial value.

- **Issue 7:** Stability analysis shows roughly a quarter of facilities move more than 3 rating deciles between quarters.

    - **Developer Response 7:** While providers may move several deciles, we want to again clarify that measure data on the NHC website is simply conveyed by displaying the provider's measure score, along with the state and national average scores for context. Scores are not displayed in terms of decile performance, so while these results are helpful to understand providers' movements within the performance distribution over time, they are not indicative of how consumers interact with these data. The high

correlation across quarters for a given provider demonstrates that the measure remains useful for the purpose of allowing consumers to distinguish systematically poor-performing and better-performing facilities.

Measure Number: 0696

Measure Title: STS CABG Composite Score

Measure Developer/Steward: Society of Thoracic Surgeons

Validity

- **Issue 1:**
  SMP members expressed concerns related to STS methodology for demonstrating empirical validity and content validity. (e.g. "An association with a different construct that is expected to be correlated with measure 0696 was not assessed… Could the developers demonstrate that the scores are associated with another related measure?")

  - **Developer Response 1:**
    With most individual measures, it is possible to find another external quality metric against which to assess validity. In the case of the STS CABG composite, we have purposely included all the major quality metrics that have been used for CABG. Thus, they are within the composite and are not available as separate external measures for validation. That is the reason we showed the correlation of the overall composite score with results for each of the domains.

- **Issue 2:**
  An SMP member suggested that the STS risk adjustment model relevant to measure 0696 needs to be updated.

  - **Developer Response 2:**
    The STS updated and published the relevant risk models in 2018;* please see "STS 2018 Adult Cardiac Surgery Risk Models: Part 2" attached. (The "Part 1" paper is also provided for additional background on the risk model updates.) Please advise if you would like us to revise 2b3.4a in the Composite Measure Testing form for this measure with the updated list of risk factors.

    *** We did not previously include the updated risk models in our measure documentation due to 2015 NQF guidance (**http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80308**) describing "decreased emphasis" in the endorsement maintenance process on updated reliability and validity (including risk adjustment) information: "**If prior testing adequate, no need for additional testing at maintenance…**" We also did not wish to**

create a mismatch between the date of our risk adjustment model and the dates of various data analyses provided in our Composite Measure Testing form.

- **Issue 3:**
  SMP members requested an update to the data used to demonstrate validity (2b1.3).

    - **Developer Response 3:**
      It is not feasible for the STS to update the analyses in 2b1.3 within the timeframe specified (by 10 AM ET on Oct. 16). Given that the SMP will not be meeting until Oct. 28-29, we will appreciate an extension to your deadline for this information, i.e. a due date closer to the SMP meeting date.

Measure Number: 1623

Measure Title: Bereaved Family Survey-Performance Measure (BFS-PM)

Measure Developer/Steward: Department of Veterans Affairs, Veteran Experience Center

---

Reliability

- **Issue 1: How Cronbach alpha used to assess reliability of individual rating not clear. Two different values of ICC1 coefficient are provided. [Reviewer 3]**

    - Developer Response 1a: **Single-item reliability estimates: Thank you for pointing this out. We present a version of reliability described by Wanous et al. (1997) for calculating single-item reliability, with a correction for attenuation due to unreliability. This formula correlates the single overall BFS overall item with the set of BFS comprising the care/communication factor. The care/communication factor was identified via principal axis factors analysis and is comprised of 7 BFS items. Based on parallel analysis, only 1-factor was extracted with factor loadings ranging from 0.64 – 0.77). The method for estimating single-item reliability uses the observed correlation between the BFS single-item and the care/communication factor, divided by the square-root of the product of overall BFS item (solving for this unknown) and Cronbach alpha for the 7-item care/communication factor:**

$$\hat{s}_{x,y} = \frac{s_{xy}}{\sqrt{s_{x,x} * s_{y,y}}}$$

Where s_{xy} is the correlation between the overall BFS item and the care/communication factor score (s_{xy} = 0.6743), s_{xx} is the Cronbach alpha of the overall BFS item (unknown), and s_{yy} is the Cronbach alpha of the care/communication factor (alpha = 0.8559). According to Nunnally (1978),

"the correlation between two such tests would be expected to equal the product of the terms in the denominator and consequently $\hat{s}_{x,y}$ ***would equal 1.0. If*** $\hat{s}_{x,y}$ ***would equal 1.0,*** s_{xy} ***would be limited only by the reliabilities of the two tests:***
$s_{xy} = \sqrt{s_{xx} * s_{yy}}$ ***(page 220)."***

The results are Estimates of Minimum Single-Item Reliability

Assume true reliability 1.00:	***Estimated reliability = 0.53***
Assume true 0.9:	***Estimated reliability = 0.66***
Assume true 0.8:	***Estimated reliability = 0.80***

Summary: Single-item reliability estimates range from 0.53 – 0.80. These are the lower boundary estimates of reliability. The true reliability could be higher but cannot be lower.

Wanous JP, Reichers AE, Hudy MJ. Overall Job Satisfaction: How Good Are Single-item Measures? *J Appl Psychol.* 1997 Apr; 82(2):247-52.

- **Developer Response 1b: ICC estimate: The correct ICC estimate, reported as a proportion is 0.04. The ICC reported as a percentage is 4.0%.**

- **Issue 2: The developer should summarize the relevant information from references instead of just listing them. In particular, they should present their risk adjustment information in a more cohesive way with more details. [Reviewer 1]**

  - **Developer Response 2:** We examined 5 variables as potential case-mix adjustors: veteran's age at the time of death (in years), number of medical comorbidities present at the time of death as defined by van Walraven and colleagues' modification of the Elixhauser score, veteran's primary diagnosis on last admission (classified into 1 of 15 clinical categories using the Agency for Healthcare Research and Quality Clinical Classification Software), relationship of veteran's next-of-kin (eg, spouse), and BFS administration mode (eg, mail). To examine relationships between case-mix variables and the BFS-PM, we constructed a set of regression models using logistic regression. Models were fit using both raw coefficients for categorical variables and standardized coefficients for continuous variables. Post-estimation tests, including Akaike information criteria (AIC) and the area under the receiver operating characteristic curve (ie, *C*-statistic), were used to assess model fit. The *C*-statistic for our adjustment model for the BFS-PM was 0.5835 with an AIC of 36128.58.

    Facility-level scores are adjusted for case mix using inverse probability weighting. Scores are weighted as follows: First, all case-mix adjustor variables are entered into a logistic regression model, predicting a response of "excellent" on the BFS-PM at

the patient level. A propensity score, or predicted probability, for an "excellent" response is derived from the results of the logistic regression. Finally, weights for the case-mix adjustment are calculated by taking the reciprocal of the propensity score and are applied all facility-level BFS outcomes. Weights are re-calculated quarterly.

- **Issue 3: Concern that only one item from the Bereaved Family Survey is used. [Reviewer 6]**
    - **Developer Response 3: The remaining items on the Bereaved Family Survey are used extensively for quality improvement purposes. We are seeking endorsement for the BFS-Performance Measure- the proportion of bereaved family members who rate the overall care as "excellent" - as only this measure is used for the purposes of facility-level comparisons and benchmarking.** However, in the future, we would consider seeking endorsement for all items on Bereaved Family Survey since we believe the individual survey items are important for quality improvement purposes.

Validity

- **Issue 1: Since this is a risk adjusted measure, validity testing should be based on adjusted scores instead of unadjusted scores [Reviewer 1]**
    - **Developer Response 1: Please see Table and text provided in section 2b1.3 for all fully adjusted BFS Overall-Performance Measure scores by end-of-life quality indicators. Table A below shows the adjusted associations between process measures and BFS-PM at the facility level. Table B displays these relationships at the patient level.**

As shown in Table A, nonresponse and patient case-mix adjusted facility-level BFS-PM scores are consistently higher when patients receive these quality indicators. Weighted linear regression analyses demonstrate statistically significant, positive associations between receipt of a quality indicator and facility-level BFS Performance Measure scores.

Table A. Adjusted Associations Between Process Measures and Facility-level BFS Overall Rating of EOL Care Across Two Modes of Administration (Telephone and Mailed Surveys)* for FY10-FY17

Process Measure	Facility-Level PM Score with (Yes) and without (No) Receipt of Process Measure		*β* coefficient	95% Confidence Interval	P-value
	YES	NO			
Telephone Survey					
Palliative Care Consult prior to death	59	56	0.03	0.03-0.03	<0.001

Table A. Adjusted Associations Between Process Measures and Facility-level BFS Overall Rating of EOL Care Across Two Modes of Administration (Telephone and Mailed Surveys)* for FY10-FY17					
Process Measure	**Facility-Level PM Score with (Yes) and without (No) Receipt of Process Measure**		***β* coefficient**	**95% Confidence Interval**	**P-value**
	YES	**NO**			
Telephone Survey					
Death in a Hospice/Palliative Care Unit	60	57	0.03	0.03-0.03	<0.001
Chaplain Contact with Veteran or Family	58	56	0.02	0.02-0.03	<0.001
Bereavement Contact with Family	59	57	0.01	0.01-0.02	<0.001
Mailed Survey					
Palliative Care Consult prior to death	61	59	0.03	0.02-0.03	<0.001
Death in a Hospice/Palliative Care Unit	62	60	0.02	0.02-0.03	<0.001
Chaplain Contact with Veteran or Family	61	59	0.03	0.02-0.04	<0.001
Bereavement Contact with Family	62	60	0.02	0.01-0.02	<0.001

*Linear regression models [weighted by facility size]

- As shown below in Table B, nonresponse and patient case-mix adjusted patient-level BFS-PM scores are consistently higher for when patients receive these quality indicators. Logistic regression analyses demonstrate statistically significant, positive associations between receipt of a quality indicator and patient-level BFS Performance Measure scores.

Table B. Adjusted Associations Between Process Measures and Patient-level BFS Overall Rating of EOL Care Across Two Modes of Administration (Telephone and Mailed Surveys)* for FY10-FY17					
Process Measure	**Patient-Level PM Score with (Yes) and without (No) Receipt of Process Measure**		**Odds ratio**	**95% Confidence Interval**	**P-value**
	YES	**NO**			
Telephone Survey					
Palliative Care Consult prior to death	60	46	1.74	1.57-1.93	<0.001
Death in a Hospice/Palliative Care Unit	64	52	1.63	1.46-1.82	<0.001

Table B. Adjusted Associations Between Process Measures and Patient-level BFS Overall Rating of EOL Care Across Two Modes of Administration (Telephone and Mailed Surveys)* for FY10-FY17					
Process Measure	**Patient-Level PM Score with (Yes) and without (No) Receipt of Process Measure**		**Odds ratio**	**95% Confidence Interval**	**P-value**
	YES	**NO**			
Telephone Survey					
Chaplain Contact with Veteran or Family	58	48	1.48	1.32-1.65	<0.001
Bereavement Contact with Family	57	55	1.09	0.99-1.20	0.071
Mailed Survey					
Palliative Care Consult prior to death	65	51	1.74	1.65-1.82	<0.001
Death in a Hospice/Palliative Care Unit	69	56	0.02	0.02-0.03	<0.001
Chaplain Contact with Veteran or Family	62	54	1.43	1.35-1.51	<0.001
Bereavement Contact with Family	62	59	1.14	1.09-1.18	<0.001

- **Issue 2: Inadequate description of what the rating is intended to capture or measure**. **[Reviewer 3]**

    - **Developer Response 2: Given that the alignment of patient/family preferences with treatment is a cornerstone of optimal EOL care, t**he purpose of the **Bereaved Family Survey Performance Measure (**BFS-PM) is to assess families´ perceptions of the overall quality of care that Veterans received from the VA in the last month of life. More specifically, **the BFS-PM score captures the proportion of family members that rate the overall care of their deceased Veteran at end of life as "Excellent".**

- **Issue 3: Nonresponse is high and systematic. Scores are adjusted via inverse probability weighting for nonresponse, and the mean difference in scores after adjustment is – 2%. There is, however, no presentation of information on the goodness of fit of the adjustment model. [Reviewer 3]**

    - **Developer Response 3:** The model fit was evaluated via the Akaike Information Criterion (AIC) and the area under the receiver operating characteristic (ROC) curve (C-statistic). Candidate models were rank-ordered according to model fit, and the model with the best C-statistic and the lowest AIC was selected. Using the model

selection criteria described, an 18-variable model with the best C-statistic (0.65) and lowest AIC (24154.07) was identified.

- **Issue 4: Low correlation with measures of high quality EOL care suggest that it is not measuring quality of EOL care. [Reviewer 3]**

  Developer Response 4: We have established construct validity in following the general approach described by Westen et al. in the *Journal of Personality and Social Psychology* (http://nrs.harvard.edu/urn-3:HUL.InstRepos:3708469**). Researchers generally establish the construct validity of a measure by correlating it with several other measures and arguing from the pattern of correlations that the measure is associated with these variables in theoretically predictable ways. These guidelines do not prescribe a recommended magnitude of association. In Tables A and B above (and in prior published work outlined in our application), we demonstrate theoretically-predictable, positive, and significant correlations between the BFS-PM and a set of 4 EOL process measures that are also indicators of high-quality EOL care. These correlations indicate that unique aspects of high-quality EOL care are being captured by the BFS-PM and the process measures. Although these process measures are theoretical and empirical correlates of the BFS-PM, they are not meant to be a substitute for it.**

- **Issue 5: Unclear rationale for inclusion of comorbidities and diagnosis in risk adjustment [Reviewer 5]**

  - **Developer Response 5: Comorbidities and diagnoses are included in the risk adjustment model because these health factors were found to have large and statistically significant effects on BFS-PM scores in our prior work (Kutney-Lee et al, 2018,** *American Journal of Hospice & Palliative Medicine***). In this paper, that describes the development of the risk-adjustment model for the BFS-PM, we also found significant differences in facility rankings before and after adjustment for the comorbidity burden of a facility's patients. Further, we include these variables to align with the Center for Medicare and Medicaid Services' (CMS) general approach for adjustment of the CAHPS surveys, including CAHPS-Hospice (**https://www.hospicecahpssurvey.org/globalassets/hospice-cahps/scoring-and-analysis/8-29-2019-updates/cma_public_document-for-website-2018q4-final.pdf**), and also in following the recommendation of AHRQ to adjust for patient severity of illness when making facility-level quality comparisons (**https://www.ahrq.gov/talkingquality/translate/scores/adjustment-scoring.html**).**

- **Issue 6: Inadequate assessment of exclusion for answering less than 12 items on survey. [Reviewer 3]**

- **Developer Response 6: We exclude surveys that do not have 12 of the 17 items (70%) completed. To date, 0.88% of returned surveys have less than 12 items completed. This determination is made a-priori and is intended to limit the amount of missing data. Extensive data imputation methods for missing data can jeopardize observing true associations between variables of interest (Brick & Kalton, *Statistical Methods in Medical Research*, 1996).**

- **Issue 7: Lack of adequate rationale for excluding social risk characteristics from risk adjustment, such as race/ethnicity and region/rural status. The reviewer stated that our rationale for exclusion of these variables is inconsistent with other surveys, such as CAHPS. [Reviewer 5]**

  - **Developer Response 7:** Although our prior work has shown that race/ethnicity and region are associated with bereaved family assessments of quality of care, we choose not to include in our risk adjustment models for the BFS-PM to more closely align with CAHPS procedures. Per CAHPS risk-adjustment methodology for facility-level comparisons of performance scores (https://www.hcahpsonline.org/globalassets/hcahps/mode-patient-mix-adjustment/october_2019_pma_web_document.pdf), race/ethnicity is NOT included as an adjustor. Further, our rationale for exclusion of these sociodemographic characteristics is also in alignment with AHRQ's recommendations for risk-adjusting performance scores for comparison purposes. AHRQ states that adjustment for such sociodemographic characteristics may essentially conceal unacceptable disparities in care. (https://www.ahrq.gov/talkingquality/translate/scores/adjustment-scoring.html).

- **Issue 8: Validity testing methods inadequate – face validity only. [Reviewer 4]**

  - **Developer Response 8: We provide extensive data regarding the predictive, construct, and discriminant validity of the measure in sections 2b1.2. and 2b1.3 in the application.**

- **Issue 9: What is the explanation for more people whose relative dies in a "low complexity" facility more likely to respond to the questionnaire instrument and how might this impact the national comparisons? How can this difference be minimized? [Reviewer 6]**

- **Developer Response 9: In the VA healthcare system, "low complexity" facilities are generally comprised of community living centers (i.e. VA nursing homes) that may have inpatient hospice units. These settings generally have the highest response rates to the survey and are also more likely to have higher scores on the BFS-PM. To account for these differences when making facility-level comparisons, we include facility complexity level as a variable in our risk adjustment model.**

- **Issue 10: Unclear what the facility level scores represent. Are they averages? Percents? I think I found some of the information in the text to indicate it was a percent but this should have been in the table. [Reviewer 6]**

  - **Developer Response 10: The facility-level score is the proportion of family members of deceased Veterans that rated overall end-of-life care as "Excellent". All 146 facility scores are then averaged into one large national mean, weighted by the number of completed surveys in each facility.**

- **Issue 11: I would have liked to have seen additional information about how the survey was developed and pilot tested. Was psychometric analysis completed? What was the rationale for including items and then not using them? What was the scoring method selected? [Reviewer 6]**

  - **Developer Response 11: We have published our methods for survey development, pilot testing, and psychometric analyses which can be found here:**

Finlay, E., Shreve, S., & Casarett, D. (2008). Nationwide veterans affairs quality measure for cancer: the family assessment of treatment at end of life. *Journal of Clinical Oncology, 26*(23), 3838-3844. doi: 10.1200/JCO.2008.16.8534.

Casarett, D., Pickard, A., Bailey, F. A., Ritchie, C., Furman, C., Rosenfeld, K., ... & Shea, J. A. (2008). Important aspects of end-of-life care among veterans: implications for measurement and quality improvement. *Journal of Pain and Symptom Management, 35*(2), 115-125. doi:10.1016/j.jpainsymman.2007.03.008.

Casarett, D., Shreve, S., Luhrs, C., Lorenz, K., Smith, D., De Sousa, M., & Richardson, D. (2010). Measuring families' perceptions of care across a health care system: preliminary experience with the Family Assessment of Treatment at End of Life Short form (FATE-S). *Journal of Pain and Symptom Management, 40*(6), 801-809. doi: 10.1016/j.jpainsymman.2010.03.019.

In prior work, we have analyzed differences in using a global item score (i.e, the BFS-PM) versus a composite score which included all BFS items. The goal was to define families' priorities for various aspects of end-of-life care, and to determine whether scores that

reflect these priorities alter facilities' quality rankings. Weights were homogeneous across patient subgroups, and there were no significant changes in facilities' quality rankings when weights were used. There appears to be wide variation in the importance that families place on several aspects of end-of-life care. However, the use of weights to account for families' priorities is not likely to alter a facility's quality score. Those results can be found in here:

Smith, D., Caragian, N., Kazlo, E., Bernstein, J., Richardson, D., & Casarett, D. (2011). Can we make reports of end-of-life care quality more consumer-focused? Results of a nationwide quality measurement program. *Journal of Palliative Medicine, 14*(3), 301-307. doi: 10.1089/jpm.2010.0321.

However, in the future, we would consider seeking endorsement for all items on Bereaved Family Survey since we believe the individual survey items are important for quality improvement purposes.

- **Issue 12: There seem to be an English and Spanish version for the instrument. Was cross-cultural validation done? Unless I missed it, I do not see that information in the application. [Reviewer 6]**

    - **Developer Response 12: We have not conducted a cross-cultural validation because most (99%) of our Spanish data collection occurs in only one facility in Puerto Rico. Therefore, it is empirically challenging to disentangle whether our findings are due to language or facility differences.**

Other General Comments

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

General Comment: We would like to share with the committee that the Bereaved Family Survey is currently being used outside of the VA, and is being fielded by the Stanford, Duke, UCLA, and Kaiser Permanente health systems. This not only speaks to the strengths and distinctiveness of the survey, but also presents unprecedented opportunities to compare quality of end-of-life care in VA and non-VA settings using the BFS-PM.

Measure Number: 2456

Measure Title: Medication Reconciliation: Number of Unintentional Medication Discrepancies per Medication Per Patient

Measure Developer/Steward: Brigham and Women´s Hospital

Validity

- **Issue 1:** Efforts need to be made to improve training so that the kappa measure of inter-rater agreement increases over time. Need to examine sources of disagreement and apply lessons learned to training materials. Request for plans for ongoing testing of interrater reliability. Concern about currently reported level of agreement (77%).

    - **Developer Response 1:** We have worked closely with Leapfrog to develop an entire suite of training materials and webinars such that the current kappa is likely higher than when originally assessed. If desired, as part of the NQF measure, we can require that a sample of cases be over-read by a second reviewer at each site to look for disagreement and then to report on any disagreement and how it was resolved. We can then use that information to continue to improve our training materials over time.

- **Issue 2:** Concern about selection of patients to be assessed (e.g., excluding patients who decline to talk to a pharmacist or who are unavailable to be seen), concern of selection bias by SES, "difficult patients," etc. Is there really no alternative to a live interview to getting a medication history?

    - **Developer Response 2:** This is a valid concern. Fortunately, in our studies we have found that very few (less than 10%) of patients decline to talk with a study pharmacist or are unavailable to be seen. We emphasize in our training materials that once a patient has been sampled for measurement, every effort should be made to take a gold standard medication history on that patient before they are discharged (i.e., minimizing the number that are unavailable). We believe this approach has been effective in minimizing excuses to not get data on difficult patients. If desired, we could ask sites to collect basic demographic data on sampled patients who participate and those who do not in order to measure any selection bias, but we are very concerned about creating additional burden on sites. Lastly, while it would be convenient if there were an alternative to determining the gold standard medication history, there really is no other way to do this other than a live interview (or a phone interview of a caregiver if they are responsible for the patient's medications, which is part of the protocol).

- **Issue 3:** Validity is only as good as the training materials. Substantive committee needs to review training materials. There is no way to tell if the med rec (even if it is a gold standard) produces correct data.

    - **Developer Response 3:** We would be happy to provide the training materials to NQF so they can be reviewed.

- **Issue 4:** Critical element validity testing could not be located. Concern that only documented face validity. Rationale for using face validity (alone) is somewhat weak and not backed up by

experts in the field. Request for concurrent or predictive validity. Concern that results shown in 2b.1.1 are somewhat inconsistent.

- **Developer Response 4:** We apologize for not making this clear. In the NQF Testing Attachment 2b.1.1., we provide evidence of empirical validity testing but placed it in the face validity section. Specifically, we provide empirical evidence that hospitals that had significant improvement in their medication discrepancy rates (the critical element of the proposed measure) from the beginning to the end of the study had a greater improvement in the proportion of patients who received patient-level medication reconciliation interventions (such as a "best possible medication history" in the emergency department) than those hospitals that did not see improvement in their discrepancy rates. The results are not perfect, but they do provide evidence that quality improvement measures can improve this outcome and that it can distinguish high-performing from low-performing hospitals in terms of their ability to improve.
- If desired, we have two other options for measuring validity
  - Correlate discrepancy rates with some of the other NQF-endorsed medication reconciliation measures. Our main concern with doing this is that all of the other measures look at process and not outcome, and as we have discussed, it is very possible for the process to look good (e.g., "check a box" that medication reconciliation has been completed) without improving the actual quality of medication reconciliation. In fact, there are times where making the process measure look good might actually be counter-productive and could make the quality go down. For example, a provider stating that a high-quality medication history has been taken when in fact it hasn't been, thus impeding downstream efforts to fix the history on that patient. This is one reason why our measure is so important to the field.
  - Correlate improvements in discrepancy rates with improvements in "harder" patient outcomes such as hospital length of stay or readmission rates. We are in fact doing these analyses in the MARQUIS2 study, but the results are not completed. The main problem with these analyses is that many factors influence these harder outcomes, and so they are not as amenable to change, thus running the risk of a negative result.
- While it is true that no TEP has officially concluded that our measure provides consistently valid scores, the MARQUIS co-investigators include some noted experts in the field, including Sunil Kripalani and Mark Williams, and the results of the first MARQUIS study, which used this measure as its primary outcome, was recently published in BMJ Quality and Safety and was personally reviewed by Kaveh Shojania, its editor-in-chief and himself an expert in medication reconciliation.

- **Issue 5:** Lack of clarity of whether the metric has been changed to be the number of medication discrepancies per medication per patient (2b3.2). Request for an example.

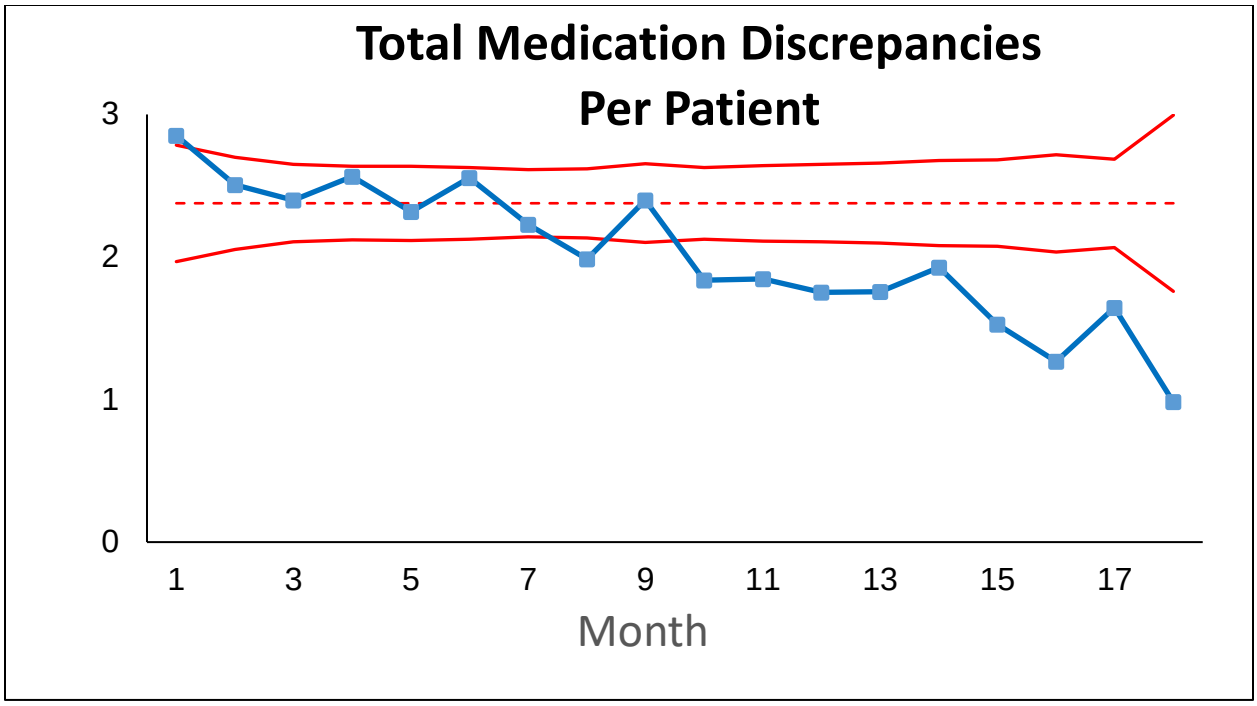
- **Developer Response 5:** We apologize for the lack of clarity on this issue. Yes, the metric has been changed to be the number of discrepancies per medication per patient. This adjusts for the number of medications each patient is on (each of which is an opportunity for error) and matches the Leapfrog measure.
- We provide an example of how to do this calculation in the Town Hall slide deck we provided last time (see slides 33-46). In that example, there are 4 gold standard medications and 1 unintentionally ordered additional medication; thus, the denominator is 5. There are 5 admission or discharge orders with unintentional discrepancies in the gold standard medications and 2 admission or discharge orders with unintentional discrepancies in the additionally ordered medication; thus the numerator is 7. So in this case, the number of medication discrepancies per medication per patient is 7/5 or 1.4. The maximum number of discrepancies per medication per patient is 2, because every medication can be ordered incorrectly at both admission and discharge.

- **Issue 6:** Concern about using a t-test in section 2b4.1 for count data.

    - **Developer Response 6:** We want to clarify that this approach was only used for the power calculations. There are ways to incorporate Poisson regression into power calculations, but they require simulation and require several assumptions which we did not feel comfortable making. So this was the best alternative. The analyses of outcomes, e.g., in the MARQUIS studies, use multivariable Poisson regression.

- **Issue 7:** On page 7 of the testing attachment, concern that the rate of discrepancies went up at the end – do providers grow indifferent over time?

    - **Developer Response 7:** This is a great point. Sites definitely need to focus on sustainability after the novelty of the intervention has worn off. Ongoing measurement should make sites aware of any slippage and provoke them to take action as necessary. We should note that in the MARQUIS2 study, where we paid a lot of attention to sustainability, discrepancy rates continued to fall across the 18 sites throughout the study period. See below for the statistical process control chart of those results (unpublished data under review at JAMA, please do not share).



- **Issue 8:** Please specify the number of cases required for this measure and specify a sampling scheme.
    - **Developer Response 8:** This is a valid point. In the MARQUIS studies, we collected data on approximately 22 patients per month (i.e., average of one per weekday). Leapfrog currently requires 20 patients per quarter, and we do worry that the number is too small (i.e., too much variation based on sampling). We would be open to specifying a number, but I am also concerned about specifying a number larger than Leapfrog's current requirements. Ideally, I can convince Leapfrog to increase its requirements, and then we can make a recommendation here that is concordant.
    - We do specify a sampling scheme in the Leapfrog documents; we apologize if it was not as clear in the NQF documents. Here they are:

"Hospitals are required to sample at least 20 patients per quarter (any consecutive three months). Patients who were discharged or expired before the gold standard history could be obtained **should be excluded from the sample**. Hospitals can use the 'Sampling' tab of this workbook to obtain the 20-patient sample. To use this tab, first scroll down to the date of admission. The columns to the right of the date contain a string of numbers which represent the patients to include in your sample based on the order of admittance for that day. For example, if you were to sample patients who were admitted on April 1st, then you would scroll down to that date and see the following numbers: "6, 13, 5, 1..." This means that the first patient to sample would be the 6th person that was admitted on April 1st. The next patient to sample would be the 13th admitted on April 1st, and so on. Note that you do not need to sample all 20 patients from one day in the quarter. This tab is designed so that you can sample a handful of patients on various days over the course of the quarter."

Measure Number: 3492

Measure Title: Acute Care Use due to Opioid Overdose
Measure Developer/Steward: Center for Outcomes Research and Evaluation (CORE)

Reliability

- **Issue 1: Reviewers requested additional detail about measure specifications.** Specifically, reviewers requested clarification about the measure denominator, including whether the denominator includes Medicare fee-for-service only and how periods of unenrollment are handled. Reviewers also requested clarification about whether the numerator and denominator are derived from the same geographic areas.

    - **Developer Response 1:** The denominator includes all adults 18 and older who are enrolled in Medicare fee-for-service during a one-year measurement period. There is no minimum period of enrollment. The denominator is measured in person-years and only includes periods of enrollment. Thus, periods of unenrollment do not contribute to the denominator and people would also not be captured in the numerator during periods of unenrollment.
    - Individuals must reside within the measured geography to be included in the denominator. For example, if we are applying the measure to the state of Maryland, only Medicare enrollees who reside in Maryland are included.
    - All events in the numerator occur among people included in the denominator and all people in the denominator could contribute to the numerator.
    - People enrolled in Medicare Advantage plans are not included.
    - People who are dually eligible for Medicare and Medicaid are included. Because Medicare is usually the primary payer, emergency department visits are visible in Medicare claims.
    - Both numerator and denominator events are attributed to a geography based on the individual's place of residence. For example, if an individual resides in Maryland but has an overdose event in Washington DC, the person would be counted in both Maryland's denominator and numerator. The rationale here is that overdose events reflect the quality of care, including availability of treatment for opioid use disorder, in the region where they live, not necessarily where they have an overdose or where they are transported during an overdose.
    - We considered specifying the measure with ED visits as the denominator, as one reviewer suggested. However, ED visits are dependent on the baseline health of a population, which may vary from place to place, particularly if the age of the beneficiary population differs from place to place or changes over time. We felt that a denominator that captures the population size is more appropriate.

Issue 2: Reliability Testing: Most reviewers felt that the reliability testing was appropriate and indicated acceptable reliability. However, some reviewers asked for additional detail beyond the range of reliability scores.

- **Developer Response 2:** We now report the Adams reliability score for each state in our sample and each county in Maryland for 2017:

Unit reliability testing for states and counties, sorted by reliability

State	Adams Reliability	MD County	Adams Reliability
CA	0.99762	Montgomery	0.99526
TX	0.9972	Prince George's	0.98714
FL	0.99597	Frederick	0.97785
IL	0.99504	Anne Arundel	0.97707
NC	0.99206	Baltimore	0.97515
GA	0.99116	Howard	0.97417
AZ	0.99098	Harford	0.9581
IA	0.99068	Talbot	0.92825
MI	0.98827	Calvert	0.9254
WI	0.98825	Carroll	0.92393
IN	0.98751	Worcester	0.9185
MO	0.98709	Washington	0.91595
TN	0.98688	Baltimore City	0.91434
MD	0.98347	Charles	0.91263
NE	0.98331	Allegany	0.88907
OR	0.98216	Saint Mary's	0.88241
KS	0.98103	Cecil	0.87122
KY	0.98008	Wicomico	0.85853
MN	0.97376	Kent	0.84559
SD	0.96798	Queen Anne's	0.8454
NV	0.96643	Garrett	0.8123
ME	0.96342	Dorchester	0.74249
MT	0.96303	Caroline	0.73586
ND	0.9543	Somerset	0.60165
WY	0.9224	--	--

- We also report results from split sample testing. For split sample testing, we randomly split the patient-level data into two halves, calculated the measure for each state and county in our sample, and compared measure results for each measured entity (state or county) using a correlation coefficient. As noted, the correlation between split samples for states was 0.94 and for counties was 0.87.

Split-sample reliability

State	State Measure Results Sample 1	State Measure Results sample 2	County	County Results Sample 1	County Results Sample 2
AZ	0.852	0.875	Allegany	2.128	0.871
CA	1.032	0.993	Anne Arundel	1.598	1.132
FL	1.272	1.234	Baltimore	2.793	2.104
GA	1.173	1.155	Baltimore City	6.040	6.250
IA	0.586	0.548	Calvert	1.144	0.562
IL	1.017	0.973	Caroline	1.614	1.665
IN	1.370	1.389	Carroll	2.310	1.243
KS	1.076	0.971	Cecil	1.539	2.221
KY	1.591	1.601	Charles	1.449	1.333
MD	1.920	1.682	Dorchester	1.130	2.547
ME	0.873	1.302	Frederick	0.587	0.688
MI	1.831	1.852	Garrett	0.686	1.344
MN	1.469	1.284	Harford	1.719	0.938
MO	1.329	1.272	Howard	0.974	0.568
MT	0.881	0.740	Kent	0.999	0.656
NC	1.213	1.327	Montgomery	0.440	0.529
ND	0.639	0.555	Prince George's	1.103	0.872
NE	0.696	0.500	Queen Anne's	1.485	0.737
NV	1.334	1.442	Saint Mary's	1.452	1.577
OR	0.979	1.012	Somerset	2.375	1.856
SD	0.600	0.463	Talbot	0.390	0.791
TN	1.352	1.356	Washington	1.818	1.814
TX	0.866	0.882	Wicomico	2.176	2.155
WI	0.958	1.005	Worcester	1.135	0.720
WY	1.077	0.946	--	--	--

- **Issue 3: Meaningful differences:** Reviewers asked for additional detail about testing for differences, including statistical tests used. Reviewers also asked about how we might track meaningful differences if the measure were used to compare entities to one another or track entities over time.

    - **Developer Response 3:** We presented two sets of meaningful differences testing. First, we evaluated whether entities (counties or states) differed from the mean using a one sample t-test and we considered a p-value of <0.05 to constitute a meaningful difference from the mean. As described in our

submission forms, 12 states had below average rates, 10 states were above average, and 3 were no different from average. Among counties, 2 were above average and 9 were below average. In this context, “above average” indicates a higher rate of overdose and worse performance.

Second, we evaluated changes in performance over time within entities from 2017 to 2018. Here we used a generalized linear model with a Poisson distribution and a population offset. We fit one model per entity with time (year) as the main effect. A p-value <0.05 for year suggested differences in performance within an entity over time. Using this method, we observed that 19/25 states had a statistically significant change in measure performance from 2017 to 2018. Among counties, we found that 3 out of 24 counties had a statistically significant change in performance between 2017 and 2018.

- **Issue 4: Risk adjustment:** Reviewers generally agreed with the rationale that risk adjustment would mask the very disparities this measure is trying to capture. One reviewer asked whether differences in Medicare Part A and B uptake might introduce bias.
    - **Developer Response 4:** We agree that varied enrollment patterns may influence the measure results. However, the vast majority of FFS beneficiaries are enrolled in both Parts A and B (93%) with about 6% enrolled in part A only and 1% enrolled in Part B only. Given that this is truly a minority of Medicare enrollees, we do not feel that this is likely to significantly bias results.

Validity

- **Issue 1: Face validity assessment:** Reviewers commented that a larger TEP, national, multi-stakeholder technical expert panel would be preferable for assessing face validity than from a small panel of Yale faculty.

    - **Developer Response 1:** Our main argument for the validity of the measure was not based on expert review of the face validity, but rather on the literature that supports the validity of the measure. A number of studies compare opioid diagnostic codes used in this measure to a gold standard, chart review. These studies suggest that diagnostic codes indicating opioid overdose are highly specific for opioid overdose with reasonable positive predictive value. In addition, we performed extensive empirical testing, comparing the measure to two other measures of opioid overdose at the state level (an AHRQ measure, and opioid overdose rates, both reported at the state level).

      To supplement this, we convened a panel of clinical experts to assess the face validity of the measure. We asked panel members to rank the measure on a Likert scale from 1-5, with 5 indicating highest face validity. All 5 members rated the measure a 4. However, this panel was never intended to be formal TEP but rather a way of using local expertise to assess and improve the measure. Thus,

although there are limitations to using a local panel, our argument for the validity of this measure has never rested on its face validity.

- **Issue 2: Empirical validity testing:** Reviewers requested additional results from the comparison between the proposed measure and the AHRQ measure. They also noted that comparing the measure to an all payer population may be problematic. One reviewer noted that AHRQ measure rates in an all-payer population were considerably higher than rates from the measure under consideration.

    - **Developer Response 2:** We believe that comparing our measure results to an independently developed measure is valuable. Further, we believe a high correlation in results even with a different outcome definition and a different population (Medicare vs all-payer) strengthens our argument that the measure captures the underlying conditions driving the opioid epidemic and is not simply reflecting idiosyncrasies of opioid overdose in the Medicare population.

      We did observe higher outcome rates in the AHRQ measure, but this is likely for two reasons. First, the AHRQ measure population is an all-payer population which may have higher rates of overdose. Second, the AHRQ measure has a much broader outcome definition and captures opioid related hospitalizations that may not be overdose per se, but could be other adverse events or conditions related to opioid use. Despite different absolute values, we did observe strong correlations between the proposed measure and the AHRQ measure (r=0.74) and between the proposed measure and opioid overdose death rates (r=0.74).

      We have provided the measure rate, AHRQ measure rate, and opioid overdose death rates at the state level:

State	Measure outcome rate (per 1000 person-years)	AHRQ Measure outcome rate (per 1000 population)	Opioid Overdose Death Rate (per 100,000 population)
MD	1.801	9.607	32.2
KY	1.596	8.413	27.9
NV	1.388	5.359	13.3
IN	1.379	5.828	18.8
MN	1.377	5.515	7.8
TN	1.354	7.103	19.3
MO	1.300	5.915	16.5
NC	1.270	5.870	19.8
FL	1.253	6.134	16.3
GA	1.164	2.552	9.7

State	Measure outcome rate (per 1000 person-years)	AHRQ Measure outcome rate (per 1000 population)	Opioid Overdose Death Rate (per 100,000 population)
ME	1.087	7.193	29.9
KS	1.024	3.192	5.1
CA	1.013	3.943	5.3
WY	1.011	2.589	8.7
OR	0.996	6.394	8.1
IL	0.995	6.081	17.2
WI	0.981	5.008	16.9
TX	0.874	2.278	5.1
AZ	0.863	5.869	13.5
MT	0.810	4.710	3.6
NE	0.598	2.493	3.1
ND	0.597	4.498	4.8
IA	0.567	2.560	6.9
SD	0.532	2.477	4

- **Issue 3: Validity as a quality measure:** One reviewer questioned whether measuring opioid overdose is a measure of the quality of a care in a population.

    - **Developer Response 3:** We agree that the rate of opioid overdose resulting in emergency department use among a geographically-defined is not a traditional process quality measure. It is not intended to be so; this is an outcome measure intended to reflect quality of care throughout the continuum, from prescribing patterns, to early identification and treatment of opioid use disorder, and risk reduction among those with opioid dependence. Multiple healthcare providers have the opportunity to intervene to reduce opioid overdose events. There is strong evidence in the scientific literature to support specific interventions in the health care domain that can reduce opioid overdose. Specifically, medication assisted treatment (methadone, buprenorphine) has been shown to reduce opioid use and lower the risk of death due to opioid overdose. Other evidence-based approaches, including reducing opioid prescribing, implementing screening, brief intervention, and referral to treatment (known as SBIRT), and distribution of naloxone also reducing harm from opioid use. Thus, health systems that provide accessible and coordinated treatment for opioid use disorder may be able to reduce overdose rates. Further, we anticipate this measure being used in a context in which health care entities are responsible for the health outcomes of geographically-defined populations, such as in the Maryland Total Cost of Care Model. Measuring opioid overdose on the population level can provide important insights about how measured entities are

addressing this pressing problem and whether entities are improving over time compared to past performance.

Other General Comments

Reviewer #4 expressed concern that this measure is really an epidemiology measure and not among the kinds of quality measures that NQF does or should evaluate and endorse. Historically, though, NQF has evaluated and endorsed measures with similar specifications. For example, NQF 2020 is a measure of current adult smoking at the state level. In addition, NQF has shown substantial interest in population health, including the development and publication of the NQF Population Health Framework. Therefore, we feel it is well within NQF's purview, interest, and expertise to evaluate measures such as the one we have put forth.

Measure Number: 3528

Measure Title: CDC and VON Late Onset Sepsis and Meningitis in Very Low Birthweight Neonates

Measure Developer/Steward: CDC and VON, Steward Daniel Pollock

---

Reliability

Issue 1: Difficult to assess reliability because measure specifications cover three different measures

- **Developer Response 1:**
  The specifications for the proposed measure include numerator and denominator details for two types of neonatal infections, namely late onset sepsis (LOS) and meningitis (MEN), for which measure data will be analyzed and summarized using several different outcome statistics, namely 1) cumulative admission risk, 2) crude monthly risk, 3) survival probability, and 4) standardized infection ratio (SIR). Cumulative admission risk is the lead outcome that reflects the risk of acquiring LOS or MEN for any eligible neonate during their admission to an eligible neonatal unit. Crude monthly risk reflects the simplest raw percent of neonates that have acquired LOS or MEN in a given unit and month. Survival probability reflects the chance that any eligible neonate will remain free of either LOS or MEN, respectively. This measure can be used to produce survival plots often referred to as Kaplan-Meier plots that help neonatology staff to better understand LOS or MEN event risk that incorporates the duration of eligibility among all neonatal patient lengths of stay. Cumulative admission risk, crude monthly risk, and survival probability summary statistics will be calculated at the neonatal patient location level, more specifically by level, II/III, III, or IV NICU. Analysis by these levels of care will serve as means of risk stratification. The fourth outcome called an SIR reflects a summary measure of the cumulative admission risk and is a ratio of observed to predicted infections. SIRs will be provided for both LOS and MEN events and be calculated at the NICU level as well as summarized at the hospital level among all eligible neonatal patient locations. This SIR is similar to those SIRs calculated for other healthcare-associated infections such as procedure-associated Surgical Site Infections (see

https://www.cdc.gov/nhsn/pdfs/ps-analysis-resources/nhsn-sir-guide.pdf). Because SIRs have lower precision for NICUs with few predicted events relative to the number of admissions, i.e., low reliability, Bayesian statistical techniques are used to derive the final reliability adjusted SIR.

Issue 2: Additional background is needed on data source and data collection methods. More details needed on the online calculator.

- **Developer Response 2:** The online calculator is an executable, algorithm-based, decision-making tool designed to enable automated determinations of whether individual patients (anonymized) meet the NHSN surveillance protocol criteria for reportable LOS or MEN events. The underlying rationale for the tool is that LOS and MEN determinations can be automated by applying executable algorithms against healthcare data that are ubiquitously available in electronic form, from admission/discharge/transfer, laboratory, antimicrobial administration, and electronic health record systems. When surveillance criteria are met, required data for each instance of LOS or MEN are assembled and delivered electronically to NHSN. This electronic supply chain obviates the need for manual data collection and processing. The calculator was developed by NHSN with the goal of enabling vendor implementers to replicate its logic in their implementations. In effect, the calculator serves as a reference implementation for processing a set of data elements and rendering a rules-based decision concerning reportability. Establishing the calculator's reliability by comparing its performance to expert review of candidate LOS and MEN cases is centrally important to assuring the electronic supply chain produces results that are identical or virtually identical to more labor intensive and costly human expert reviews of source data elements. Our method of reliability testing is designed to demonstrate that the measure data elements are repeatable, producing the same results a high proportion of time when assessed in the same population in the same time period, as per the description of reliability that NQF includes in the MIF (2a2).

Data elements embedded in the algorithm include:

	Numerator	Denominator
Date of birth	√	√
Date of NICU admission, transfer, or discharge	√	√
Location of birth (inborn or outborn)	√	√
Birth weight	√	√
Gestational age	√	√
Dates of all positive blood or cerebrospinal fluid cultures	√	
All bacterial pathogens, common commensals, and fungal organisms on the NHSN organisms list that were identified in diagnostic microbiologic testing and for which test results are	√	

	Numerator	Denominator
available in the hospitals laboratory information system (LIS) and/or electronic health record system (EHRs)		
Dates of administration and name of each intravenously administered antimicrobial agent.	√	

Issue 3: LOS and MEN are not operationally defined

- **Developer Response 3:** Flowcharts for denominator and event determination, and outlining the requirements for each event, are provided below.

Issue 4: Clearer explanation need of methods used for reliability testing

- **Developer Response 4:**
  - Hospital records of 300 infant patients were included in reliability testing: records of 100 hospital stays from each of three facilities. The patients included from Facility 1 and 2 had a hospital stay for calendar year 2017 and had been admitted to the NICU during their stay. Patients included from Facility 3 were selected from hospital stays in either 2016 or 2017, which took into account the relatively lower patient volume of eligible infants in Facility 3.
  - Hospital record abstractors review the records of eligible infant patients and collected date of NICU admission, date of discharge, location of birth (inborn or outborn), birth weight and gestational age. Abstractors also reviewed diagnostic microbiologic test results and recorded the dates of all positive blood or cerebrospinal fluid cultures, all bacterial pathogens, common commensals, and fungal organisms included in NHSN's list of reportable pathogens along with the name and administration date for each intravenously administered antimicrobial agent.
  - Following the manual data collection by hospital record abstractors, an epidemiologist with subject matter expertise in neonatal LOS and meningitis manually applied the NHSN LOS/MEN surveillance protocol criteria to each set of abstracted data and identified all infants that met denominator inclusion criteria and within that group all infants that met either LOS or MEN case criteria Anonymized data—without an indication of whether surveillance criteria were met—was submitted to the NHSN LOS/Meningitis Calculator, hosted at CDC, and agreement calculated between the epidemiologist's determinations and the calculator results. In the event of a discrepancy, data was reviewed with the site abstracting resource to verify the accuracy of the abstracted data, and variances were documented. The abstracted data and determinations by the epidemiologist served as the reference standard. True Positives (TP) were defined as infants who had LOS/Meningitis by the abstracted data and the calculator. False Positives (FP) were defined as infants who had LOS/Meningitis by the calculator but not the abstracted data. False Negatives (FN) were defined as infants who had LOS/Meningitis by abstracted data but not the calculator. The metrics of precision (TP/TP+FP), recall (TP/TP+FN), and Cohen's

Kappa were then assessed to determine the overall performance of the approach. For a desired Kappa coefficient of 0.85, at an alpha level of 0.05 and a beta level of 0.05, the analysis required at least 250 infants.

- The calculated precision was 100% and recall was 96%. Cohen's kappa coefficient was 0.96, as indicated in the table below. The calculator returned six false negatives due to data entry issues. The hospital chart abstractor at one hospital mis-recorded the names of several bacterial pathogens included on NHSN's list of reportable pathogens, which resulted in false negative calculator determinations.

Events		**Manual Abstract**		
		Y	N	*Totals*
Calculator	Y	134	0	134
	N	6	180	186
	Totals	140	180	**320**
	Precision (TP/TP+FP)		100%	
	Recall (TP/TP+FN)		96%	
	Cohen's kappa		0.96	

- The high precision statistics (100%) provide strong evidence that the algorithmically-based case determination that are central to the CDC/VON harmonized measure produce reliable results. Further, the high value of the Cohen's kappa coefficient (0.96), based on data collection and case determinations across three hospitals. supports measure reliability across healthcare facility settings and hence comparability of measure results.

Issue 5: Reliability testing "across organizations" may exceed what was actually studied

- **Developer Response 5:**

Reliability testing was performed across three hospitals. The profiles of the hospitals are provided in the table below:

	1	**2**	**3**
Region	Southwest	Northeast	Mid-Atlantic
Annual NICU Admissions	650	900	600
NICU Beds	65	16	36
NICU Stepdown beds	0	32	18
EHR system	Cerner	Metavision	Crib Notes
Teaching Hospital	Yes	Yes	Yes
Free standing Children's Hospital	No	No	Yes
Hospital Ownership	Non-Profit	Non-Profit	For-Profit

	1	2	3
NICU Type	Level IV; All surgeries including cardiac surgery requiring bypass	Level III; Surgeries except cardiac surgery requiring bypass	Level III; Surgeries except cardiac surgery requiring bypass

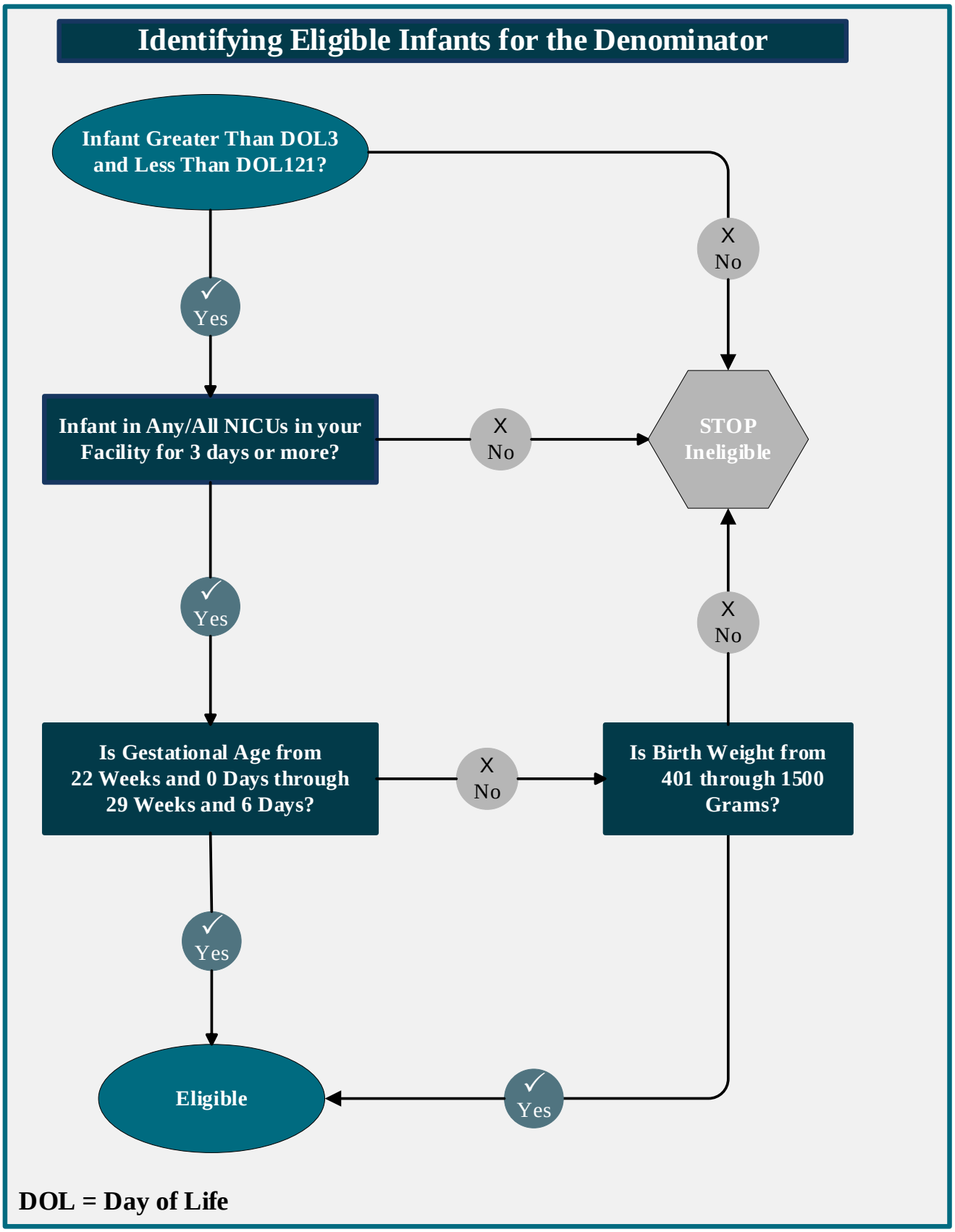
Issue 6: Extent to which and rationale for using previous measure NQF #304 to justify validity and reliability

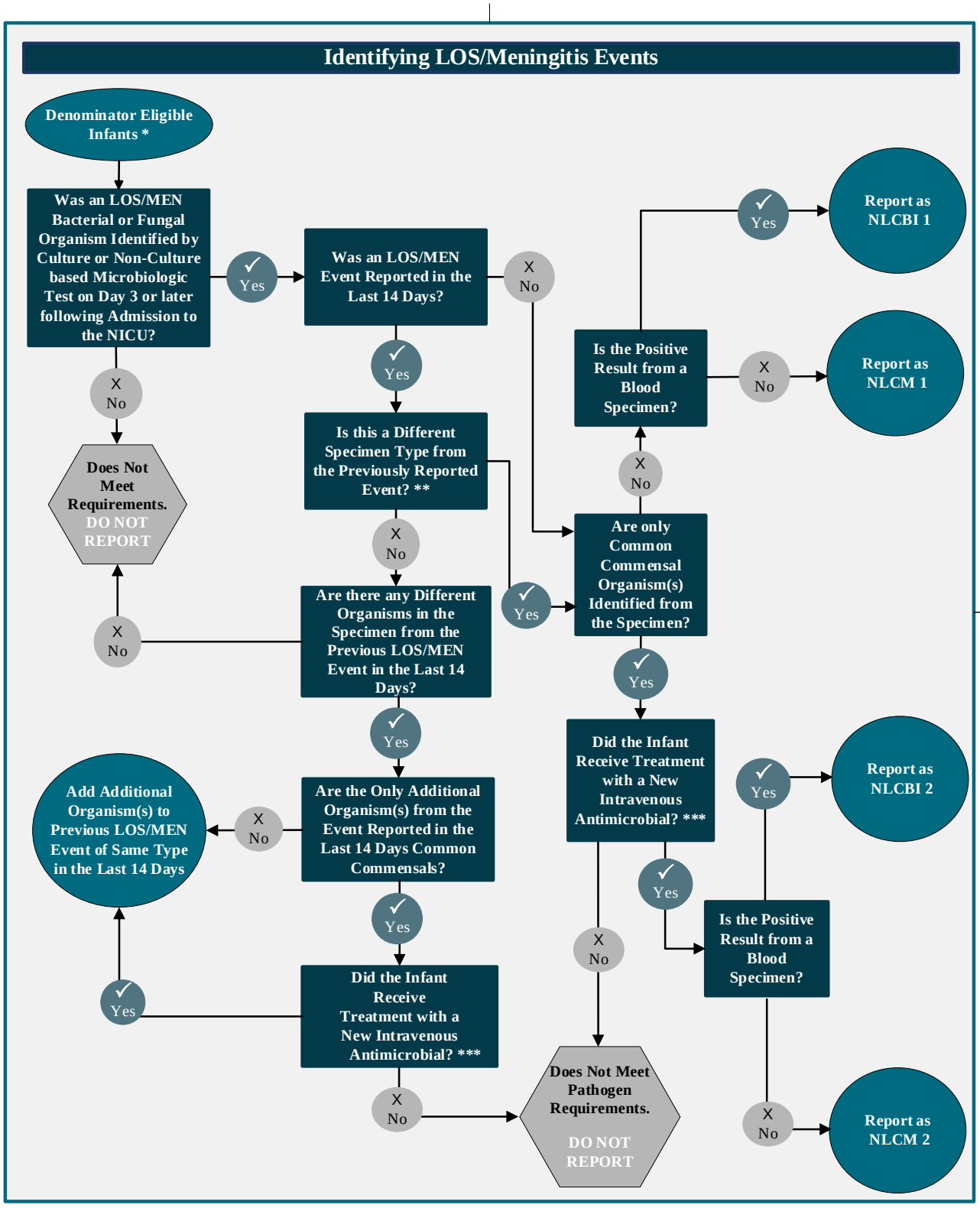
- **Developer Response 6:**

NQF#304 was not used directly to assess reliability and validity. However, the validity demonstrated by VON for the infection definitions used in NQF #304 is relevant to the NHSN measure because of the close similarities of VON and NHSN definitions for LOS.

Other General Comments

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


Identifying Eligible Infants for the Denominator
Infant Greater Than DOL3 and Less Than DOL121?
X No
✓ Yes
Infant in Any/All NICUs in your Facility for 3 days or more?
X No
STOP Ineligible
✓ Yes
X No
Is Gestational Age from 22 Weeks and 0 Days through 29 Weeks and 6 Days?
X No
Is Birth Weight from 401 through 1500 Grams?
✓ Yes
Eligible
✓ Yes
DOL = Day of Life


Identifying LOS/Meningitis Events
Denominator Eligible Infants *
Was an LOS/MEN Bacterial or Fungal Organism Identified by Culture or Non-Culture based Microbiologic Test on Day 3 or later following Admission to the NICU?
Yes
No
Was an LOS/MEN Event Reported in the Last 14 Days?
Yes
No
Does Not Meet Requirements. DO NOT REPORT
Is this a Different Specimen Type from the Previously Reported Event? **
Yes
No
Are there any Different Organisms in the Specimen from the Previous LOS/MEN Event in the Last 14 Days?
Yes
No
Add Additional Organism(s) to Previous LOS/MEN Event of Same Type in the Last 14 Days
Are the Only Additional Organism(s) from the Event Reported in the Last 14 Days Common Commensals?
Yes
No
Did the Infant Receive Treatment with a New Intravenous Antimicrobial? ***
Yes
No
Does Not Meet Pathogen Requirements. DO NOT REPORT
Are only Common Commensal Organism(s) Identified from the Specimen?
Yes
No
Is the Positive Result from a Blood Specimen?
Yes
No
Report as NLCBI 1
Report as NLCM 1
Did the Infant Receive Treatment with a New Intravenous Antimicrobial? ***
Yes
No
Report as NLCBI 2
Is the Positive Result from a Blood Specimen?
Yes
No
Report as NLCM 2

Measure Number: 3533e

Measure Title: Hospital Harm – Severe Hyperglycemia

Measure Developer/Steward: IMPAQ International (developer) / Centers for Medicare and Medicaid Services (steward)

Reliability

- **Issue 1:** A limitation is that beta-binomial parameters were estimated from only ~6 hospitals and may not be generalizable to the entire hospital population of interest.

    - **Developer Response 1: Thank you for the thoughtful analysis and reliability rating.** We understand the value of increased sample sizes in measure testing and strive for a broader set of facilities and EHR platforms whenever possible. We, however, note that measure testing was done in compliance with NQF scientific acceptability requirements specific to eCQMs, which requires "*documentation of testing on more than one Electronic Health Record (EHR) system from more than one EHR vendor is required to establish Scientific Acceptability, indicating that the measure data elements are valid and that the measure score can be accurately calculated*" in the NQF Measure Developer Guidebook August 2018 edition (page 23). We used data from six hospitals across three different EHR systems, which represent a breadth of geographic locations, patient demographic characteristics, and hospital features, which exceeds the NQF standards presented in the Guidebook. The empirical findings demonstrated high measure reliability as confirmed by the SMP's unanimous evaluation.

Validity

- **The measure developer wishes to thank the Scientific Methods Panel for their thoughtful analysis and evaluation of validity as demonstrated by the empirical findings.**

Other General Comments

Thank you for the opportunity to respond to the preliminary analysis of the NQF Scientific Methods Panel.

Measure Number: 3534

Measure Title: 30 Day All-cause Risk Standardized Mortality Odds Ratio following Transcatheter Aortic Valve Replacement (TAVR).
Measure Developer/Steward: ACC and STS

Reliability

- **Issue 1:** Concerns with small sample (40 records across 4 facilities) for interrater reliability results (IRR) as well as only providing results on "critical data elements" and not risk variables.

    - **Developer Response 1:** Per NQF criteria we provided IRR results on "critical data elements". We are not able to assess all model variables because of competing regulatory requirements for post approval studies in the TVT Registry. We are re-evaluating adding additional risk model variables in future years.

      In response to some of these concerns, we are providing updated IRR results to include 2016 and 2017 data, and 24 (of 41) model variables that were not included in the initial testing documents (in addition to the 6 critical data elements). This additional information (across years and with additional data elements) reflects a continued high agreement rate and good understanding of data definitions and consistency between the auditors.

Table R1: **IRR results (2016 and 2017 data)**

SeqNo	Data Element	Proc. Type	Matches	Universe	Agreement Rate (IRRA)	PABAK		
						Score	Lower 95% CI	Upper 95% CI
2050	Birth Date (DOB)	All	55	55	100.0%	1.000	1.00	1.00
2060	Sex (Sex)	All	55	55	100.0%	1.000	1.00	1.00
4010	Permanent Pacemaker (Pacemaker)	All	55	55	100.0%	1.000	1.00	1.00
4020	Prior PCI (PriorPCI)	All	55	55	100.0%	1.000	1.00	1.00
4030	Prior CABG (PriorCABG)	All	55	55	100.0%	1.000	1.00	1.00

SeqNo	Data Element	Proc. Type	Matches	Universe	Agreement Rate (IRRA)	PABAK		
						Score	Lower 95% CI	Upper 95% CI
4060	Prior Aortic Valve Procedure (PriorAorticValve)	All	55	55	100.0%	1.000	1.00	1.00
4120	Prior Stroke (PriorStroke)	All	55	55	100.0%	1.000	1.00	1.00
4130	TIA (CVDTIA)	All	55	55	100.0%	1.000	1.00	1.00
4145	Peripheral Arterial Disease (PriorPAD)	All	55	55	100.0%	1.000	1.00	1.00
4165	Diabetes Mellitus (Diabetes)	All	55	55	100.0%	1.000	1.00	1.00
4175	Currently on Dialysis (CurrentDialysis)	All	55	55	100.0%	1.000	1.00	1.00
4180	Chronic Lung Disease (ChrLungD)	All	52	55	94.5%	0.891	0.73	1.00
4181	Home Oxygen (HMO2)	All	55	55	100.0%	1.000	1.00	1.00
4182	Hostile Chest (HostileChest)	All	55	55	100.0%	1.000	1.00	1.00
4185	Immunocompromise (ImmSupp)	All	54	55	98.2%	0.964	0.86	1.00
5005	Prior MI (PriorMI)	All	54	55	98.2%	0.964	0.86	1.00
5045	Porcelain Aorta (PorcelainAorta)	All	54	55	98.2%	0.964	0.86	1.00
5050	Atrial Fibrillation/Flutter (AFibFlutter)	All	55	55	100.0%	1.000	1.00	1.00
5085	Five Meter Walk Test Performed (FiveMWalkTest)	TAVR	41	43	95.3%	0.907	0.73	1.00
5169	KCCQ-12 Performed (KCCQ12_Performed)	All	54	55	98.2%	0.964	0.86	1.00
5200	Height (Height)	All	51	55	92.7%	0.855	0.67	1.00
5205	Weight (Weight)	All	50	55	90.9%	0.818	0.61	1.00

SeqNo	Data Element	Proc. Type	Matches	Universe	Agreement Rate (IRRA)	PABAK		
						Score	Lower 95% CI	Upper 95% CI
5255	Pre-Procedure Creatinine (PreProcCreat)	All	53	55	96.4%	0.927	0.79	1.00
5565	Left Ventricle Ejection Fraction (LVEF)	All	51	55	92.7%	0.855	0.67	1.00
6040	Procedure Start Date (TVTProcedureStartDate)	All	55	55	100.0%	1.000	1.00	1.00
6200	Valve Sheath Access Site (TVTAccessSite)	TAVR	43	43	100.0%	1.000	1.00	1.00
9045	Discharge Date (DCDate)	All	55	55	100.0%	1.000	1.00	1.00
9050	Discharge Status (DCStatus)	All	55	55	100.0%	1.000	1.00	1.00
	Baseline Overall Accuracy		2315	2370	**97.7%**	0.954	0.94	0.97
10010	Follow-up Status (F_Status)	All	32	32	100.0%	1.000	1.00	1.00
10020	Follow-up Date of Death (F_DeathDate)	All	0	0	N/A	N/A	N/A	N/A
	30-Day Follow-up Overall Accuracy		174	175	**99.4%**	0.989	0.96	1.00

Validity

- **Issue 1:** Concern over threats to validity due to hospitals not receiving feedback because of restrictive inclusion criteria (inclusion criteria is hospitals with ≥90% complete non-missing data for 30-day mortality status, baseline KCCQ-12 score and baseline gait speed). In addition, concerns that there may be important differences between the sites that met and didn't meet the inclusion criteria.

    - **Developer Response 1:** We understand your concern over the # sites that do not meet inclusion criteria. Here are a few comments related to that criteria.

      Clinical importance of KCCQ and gait speed: Physician leaders and model developers feel it is important to use assessment of health status (via KCCQ-12) and frailty (via 5-meter walk test) in our risk models (especially for this patient population). Documentation of baseline KCCQ is required to meet the CMS "Coverage with Evidence Determination" for TAVR, describing and monitoring symptoms, functional status and quality of life for patients with heart failure. Worse baseline KCCQ scores are associated with higher risk for mortality after TAVR. In addition, slower gait speed, which is an important marker of frailty, independently predicts risk of mortality after TAVR.

      Determination of >=90% completeness threshold: In 2016, model developers reviewed different data completeness threshold's impact on # of sites and patients included (see table V1 below). Based on a review of data completeness at different thresholds, they felt we should limit analysis and hospital feedback to sites with >=90% completeness on these variables to improve internal validity. KCCQ and gait speed were imputed to the median for patient records that had missing data. Imputation slightly penalizes sites because they don't benefit from full risk adjustment (patients with missing data may appear to be less sick than they actually are). Since 90% is the standard data quality completeness threshold for all data elements in risk models, we felt this bar of 90% was reasonable, given the expectations to perform these assessments.

      Differences between sites that were included/excluded: We understand your concerns over differences in included and exclude sites. We did provide an analysis between the two groups. As shown in the Table 2b3.2 of the testing form, there were no significant differences in teaching status, bed size, or annual TAVR procedural volume between included and excluded sites. There were few meaningful differences between patients from included and excluded sites, with

patients from included sites being less likely to be of nonwhite race or Hispanic ethnicity (6.4% vs. 11.1%, standardized difference 17%) and more likely to have a tricuspid aortic valve (94.7% vs. 87.6%, standardized difference 25%). The rate of death at 30 days was also similar between included and excluded sites (4.7%vs. 5.1%, standardized difference 2%).

Improvement: We have expected a slow improvement of the # of sites included over time from initial development (since the model reports a "rolling 3 year" timeframe, it takes a while for a site to catch up on data completeness). As documented in 2b2.3 of the testing form, there has been an improvement in the # of sites included (188 hospitals in initial development; 301 sites in the 2018q4 published outcome reports). We continue to monitor this in the future.

Figure V1:

Completeness: TVTR 30 Day Mortality/Some In-hospital KCCQ-12 Component/Some 5 Meter Walk Time

Percent of Records with Complete Data					
Site Completeness Threshold	# Sites Included	# Records Included	Complete for TVTR 30 Day Death[1]	Complete for Some In-hospital KCCQ-12[2]	Complete for Some 5m Walk Time[3]
≥ 0%	450	60770	90.0%	88.3%	82.3%
≥ 10%	435	59781	90.0%	89.4%	83.4%
≥ 20%	429	58828	90.1%	90.0%	84.5%
≥ 30%	426	58362	90.6%	90.1%	84.5%
≥ 40%	416	57452	90.9%	90.2%	85.0%
≥ 50%	397	54175	91.8%	91.5%	87.1%
≥ 60%	375	50943	92.3%	92.7%	88.7%
≥ 70%	345	45472	93.0%	93.7%	91.1%
≥ 80%	281	34747	94.4%	96.1%	94.0%
≥ 90%	188	22506	96.2%	97.8%	96.7%
≥ 100%	19	402	100.0%	100.0%	100.0%

[1]TVTR 30 Day Death is non-missing (see specs for definition)
[2]Some In-hospital KCCQ-12 (#5170-5181) component value is non-missing. There are multiple questions incorporated into the KCCQ-12
measurement. This table reports subjects in which atleast some of these questions are answered.
[3]Some 5m Walk Time (#5090, #5095, #5100) atleast one entry is non-
missing or patient is Unable to Walk (#5085 = 2)

- **Issue 2:** Concerns with providing audit results on "critical data elements" and not all risk variables.

    - **Developer Response 2:** Per NQF criteria we provided audit results on "critical data elements". We are not able to audit all model variables because of competing regulatory requirements for post approval studies in the TVT Registry and are re-evaluating adding additional risk model variables in future years.

      In response to some of these concerns, the below tables include updated audit results to include 2016 and 2017 data, as well as 24 (of 41) model variables that were not included in the initial testing documents (in addition to the 6 critical data elements).

Table V2a: TVT Registry 2016 and 2017 Audit Results: Categorical Variables – Agreement Rate and PABAK Scores

SEQNO	Data Element	Uni-verse	Year audited	PABAK Score	PABAK 95% CI	Agreement Initial Score %	Agreement Final Score %	10th Percentile -90th Percentile
2050	Birth Date (DOB)	400	2016	0.98	0.953-1.000	99.0%	99.0%	100-100
		500	2017	0.972	0.94-1.000	96.8%	96.8%	97-100
2060	Sex (Sex)	400	2016	0.97	0.937-1.000	98.5%	98.5%	90-100
		500	2017	0.980	0.96-1.00	99.0%	99.0%	98-100
4010	Permanent Pacemaker (Pacemaker)	400	2016	0.92	0.867-0.973	96.0%	96.3%	90-100
		500	2017	0.964	0.93-1.00	98.2%	98.2%	97-100
4020	Prior PCI (PriorPCI)	400	2016	0.91	0.854-0.9666	95.5%	95.8%	90-100
		500	2017	0.956	0.92-0.99	97.8%	97.8%	96-99
4030	Prior CABG (PriorCABG)	400	2016	0.99	0.970-1.000	99.5%	99.5%	100-100
		500	2017	0.984	0.96-1.00	99.2%	99.2%	98-100
4060	Prior Aortic Valve Procedure (PriorAorticValve)	400	2016	0.97	0.929-1.000	98.3%	98.3%	90-100
		500	2017	0.992	0.98-1.00	99.6%	99.6%	99-100
4120	Prior Stroke (PriorStroke)	400	2016	0.94	0.893-0.987	97.0%	97.3%	90-100
		500	2017	0.928	0.88-0.97	96.4%	96.4%	94-99
4130	TIA (CVDTIA)	400	2016	0.95	0.907-0.993	97.5%	97.5%	90-100
		500	2017	0.920	0.87-0.97	96.0%	96.0%	94-98
4145	Peripheral Arterial Disease (PriorPAD)	400	2016	0.88	0.816-0.944	94.0%	94.3%	85-100
		500	2017	0.808	0.74-0.88	90.4%	90.4%	87-94

SEQNO	Data Element	Uni-verse	Year audited	PABAK		Agreement		
				Score	95% CI	Initial Score %	Final Score %	10th Percentile -90th Percentile
4165	Diabetes Mellitus (Diabetes)	400	2016	0.96	0.914-0.996	97.8%	97.8%	90-100
		500	2017	0.928	0.88-0.97	96.4%	96.4%	94-99
4175	Currently on Dialysis (CurrentDialysis)	400	2016	0.99	0.970-1.000	99.5%	99.5%	100-100
		500	2017	0.984	0.96-1.00	99.2%	99.2%	98-100
4180	Chronic Lung Disease (ChrLungD)	400	2016	0.68	0.589-0.771	84.0%	84.8%	65-100
		500	2017	0.760	0.69-0.83	88.0%	88.0%	84-92%
4181	Home Oxygen (HMO2)	400	2016	0.92	0.867-0.973	96.0%	96.3%	90-100
		500	2017	0.956	0.92-0.99	97.8%	97.8%	96-99
4182	Hostile Chest (HostileChest)	400	2016	0.95	0.900-0.990	97.3%	97.5%	90-100
		500	2017	0.920	0.87-0.97	96.0%	96.0%	94-98
4185	Immunocompromise (ImmSupp)	400	2016	0.95	0.900-0.990	97.3%	97.3%	90-100
		500	2017	0.948	0.91-0.99	97.4%	97.4%	96-99
5005	Prior MI (PriorMI)	400	2016	0.85	0.780-0.920	92.5%	92.5%	80-100
		500	2017	0.868	0.81-0.93	93.4%	93.6%	91-96
5045	Porcelain Aorta (PorcelainAorta)	400	2016	0.96	0.922-0.998	98.0%	98.3%	90-100
		500	2017	0.980	0.96-1.00	99%	99%	98-100
5050	Atrial Fibrillation/Flutter (AFibFlutter)	400	2016	0.93	0.880-0.980	96.5%	96.8%	90-100
		500	2017	0.912	0.86-0.96	95.6%	95.6%	93-98
5085	Five Meter Walk Test Performed (FiveMWalkTest)	357	2016	0.82	0.741-0.900	91.0%	91.3%	70.7-100
		445	2017	0.717	0.63-0.80	85.8%	86.1%	82-90
5169	KCCQ-12 Performed (KCCQ12_Performed)	400	2016	0.92	0.867-0.973	96.0%	96.0%	85-100
		500	2017	0.868	0.81-0.93	93.4%	93.4%	91-96
5695	MV Insufficiency (VDInsufM)	383	2016	0.63	0.532-0.726	81.5%	81.7%	55-100
		500	2017	0.588	0.50-0.67	79.4%	79.4%	75-84
6040	Procedure Start Date (TVTProcedureStartDate)	689	2016	1.00	1.000-1.000	100.0%	100.0%	100-100
		500	2017	1.00	1.00-1.00	100.0%	100.0%	100-100
6200		357	2016	0.99	0.979-1.000	99.7%	99.7%	100-100

SEQNO	Data Element	Uni-verse	Year audited	PABAK Score	PABAK 95% CI	Agreement Initial Score %	Agreement Final Score %	Agreement 10th Percentile -90th Percentile
	Valve Sheath Access Site (TVTAccessSite)	445	2017	1.00	1.00-1.00	100.0%	100.0%	100-100
9045	Discharge Date (DCDate)	400	2016	0.97	0.937-1.000	98.5%	98.5%	90-100
		500	2017	0.960	0.93-0.99	98.0%	98.0%	96-100
9050	Discharge Status (DCStatus)	400	2016	1.00	1.000-1.000	100.0%	100.0%	100-100
		500	2017	0.996	0.98-1.00	99.8%	99.8%	99-100
10010	Follow-up Status (F_Status)	387	2016	0.77	0.689-0.856	88.6%	89.9%	70-100
		302	2017	0.980	0.95-1.00	99.0%	99.3%	98-100
10020	Follow-up Date of Death (F_DeathDate)	8	2016	0.50	0.000-1.000	75.0%	75.0%	0-100
		3	2017	N/A	N/A	33.3%	33.3%	0-100

Table V2b: TVT Registry 2016 and 2017 Audit Results: Continuous Variables – Agreement Rate and Pearson Correlation Scores

SEQNO	Field Name	Universe	Year audited	Pearson Correlation Score	Pearson Correlation Lower 95% CI – Upper 95% CI	Agreement Initial Score %	Agreement Final Score %	Agreement 10th Percentile -90th Percentile
5200	Height (Height)	400	2016	0.966	0.959-0.972	86.3%	86.5%	60-100
		500	2017	0.911	0.89-0.92	92.4%	92.6%	90-96
5205	Weight (Weight)	400	2016	0.983	0.979-0.986	75.8%	75.8%	35-100
		500	2017	0.982	0.98-0.98	79.6%	79.6%	75-84
5255	Pre-Procedure Creatinine (PreProcCreat)	400	2016	0.999	0.999-0.999	92.3%	92.3%	80-100
		500	2017	0.986	0.98-0.99	85.0%	85.0%	85-92
5565	Left Ventricle Ejection Fraction (LVEF)	398	2016	0.963	0.956-0.970	77.9%	78.6%	40-100
		500	2017	0.931	0.92-0.94	68.2%	68.2%	63-74

- **Issue 3:** Concerns of validity of 30-day follow-up date of death

    - **Developer Response 3:** The incidence of death captured post discharge up to 30 days is infrequent, making it difficult to validate in audits. To validate accuracy of 30-day mortality in the TVT Registry, we compared TVT Registry data linked CMS claims data from 2012-2015 (refer to the yellow highlighting in the table below). Across 3.5 years, 99.6% of the 29,247 patient records had no discrepancy.

Table V3: TAVR POPULATION: 30 Day Mortality - CMS vs Site Reported (Jan 2012 - Jun 2015) Among all TAVR Procedures/Lab Visits

Variable	Level	Overall (N=41582)		2012 (N=4656)		2013 (N=9104)		2014 (N=16389)		2015 Q1/Q2 (N=11433)	
Using Registry Only Data											
30 Day Death (Among non-missing)	No	34884	93.72	3901	92.53	7734	92.92	14029	93.91	9220	94.62
	Yes	2337	6.28	315	7.47	589	7.08	909	6.09	524	5.38
30 Day Death (Among entire registry)	Missing	4361	10.49	440	9.45	781	8.58	1451	8.85	1689	14.77
	No	34884	83.89	3901	83.78	7734	84.95	14029	85.60	9220	80.64
	Yes	2337	5.62	315	6.77	589	6.47	909	5.55	524	4.58
Using CMS Only Data											
30 Day Death (Among linked procedures)	No	27607	94.03	3092	92.63	6076	93.23	10867	94.27	7572	94.93
	Yes	1752	5.97	246	7.37	441	6.77	661	5.73	404	5.07
Using CMS & Registry Data											
30 Day Death Discrepancy (Among linked procedures)	No	29222	99.53	3320	99.46	6486	99.52	11476	99.55	7940	99.55
	Yes	137	0.47	18	0.54	31	0.48	52	0.45	36	0.45
30 Day Death Discrepancy: Reg Y, CMS N (Among linked procedures)	No	29334	99.91	3337	99.97	6511	99.91	11516	99.90	7970	99.92
	Yes	25	0.09	1	0.03	6	0.09	12	0.10	6	0.08
30 Day Death Discrepancy: Reg N, CMS Y (Among linked procedures)	No	29247	99.62	3321	99.49	6492	99.62	11488	99.65	7946	99.62
	Yes	112	0.38	17	0.51	25	0.38	40	0.35	30	0.38

- **Issue 4:** One panel member suggested replacing the procedure variable "access site" with a pre-procedure variable.
    - **Developer Response 4:** Currently there are no pre-procedure variables to capture pre-existing femoral artery pathology. However, we have added data elements to capture this the future version update. Once the new version is implemented and the model will be revised, we'll take this into consideration.

      O'Brien, et al[1] describes the STS and ACC's rational used when selecting care processes (including TAVR access site) to provide information about the patient's baseline clinical status:

      > *"An important principle of risk adjustment is to adjust for patient factors that are beyond the control of the entity being assessed and to avoid adjusting for factors that result from the care provided. For example, it is generally unadvisable to adjust for discretionary care processes such as intraoperative medications and procedural techniques. On the other hand, certain care processes tend to be given to patients who have a relatively serious preoperative presentation. In some cases, knowledge that a care process was delivered (e.g. preoperative intra-aortic balloon pump [IABP] or inotropes) may provide indirect information about unmeasured aspects of the patient's baseline status. For these reasons, preoperative use of a mechanical assist device and preoperative inotropes were incorporated into our risk adjustment model. Similarly, use of a non-femoral valve sheath access site was adjusted in the model. Although the decision to use a non-femoral access site is under the control of the care provider, patients receiving non-femoral access differ systematically from conventional access patients in ways that are not fully captured by other TVT data elements. For this reason, non-femoral was included as a covariate. As a result of adjusting for access site, multivariable analysis may obscure or "adjust away" some true differences in quality that are reflected in the adoption of more or less effective care processes (e.g., the decision to use femoral or alternative access). Prior to deciding to adjust for access site, a sensitivity analysis was conducted to assess the impact of adjusting versus not adjusting for this variable. The Pearson correlation between risk-adjusted mortality rates (RAMRs) calculated with (vs. without) adjustment for access site was 0.989."*
      >
      > *[1]O'Brien, S.M. et al. Variation in Hospital Risk-Adjusted Mortality Following TAVR in the U.S. A Report from the STS/ACC TVT Registry. Circulation CV Quality Outcomes, 2016; 9:560-565*

Other General Comments

-Some of our responses to validity address concerns that the SMP expressed in their evaluation of reliability.

-We've attached the published manuscript that was submitted as part of the testing document for your reference.

Measure Number: 3537

Measure Title: Intraoperative Hypotension among Non-Emergent Noncardiac Surgical Cases

Measure Developer/Steward: Developer: Mathematica. Co-stewards: Cleveland Clinic and ePreop.

We thank the panel for their thorough and thoughtful review, and for the opportunity to provide clarifications and additional analyses. In the sections below, we describe each issue raised by the panel and our response. After each issue number, we reference in parentheses the question number and reviewer number from the NQF Preliminary Analysis Form, for easy cross-referencing across documents. For example, reviewer 1's comment on question 2 is referenced as "(Q2 R1)". We also appreciate your positive comments, although we do not individually list or react to them below.

Key additions to note:

- New tables showing reliability of the risk-adjusted measure (O:E ratio) by clinician case volume. See p.6-7.
- New graph showing validity of the risk-adjusted measure score: incidence of adverse patient outcomes by clinician O:E ratio quintile. See p.20.

Reliability

- **Issue 1 (Q2 R1):** Specifications. Step 5 may be missing a step in the calculation of the clinician level risk-adjusted score. Should each probability for each case be transformed into a binary 0/1 score before summing the 'probabilities' per clinician?

    - **Developer Response 1:** To estimate the expected number of cases of IOH per clinician, we sum up the predicted probabilities (ranging from 0 to 1) for each case attributed to that clinician. For example, if the clinician has 5 cases with predicted probabilities of 0.1, 0.6, 0.4, 0.8, and 0.2, her expected number of cases of IOH overall is 2.1. That sum becomes the "E" used to calculate the clinician-level O:E ratio. It is not necessary to transform each case's predicted probability into a binary score.

- **Issue 2 (Q2 R1):** Specifications. Step 7, the optional transformation of O:E ratio to pseudo-percentage, is confusing.

    - **Developer Response 2:** Thank you for the suggestions on how to make this optional step clearer for users. We plan to update the specifications as needed, and will plan to remove the word "percentage" and provide an example in Step 7, as the reviewer suggested.

- **Issue 3 (Q2 R2):** Specifications. Clarify whether a hierarchical model was used, as the measured entity is the individual clinician.

- **Developer Response 3:** We believe the reviewer may be referring to our approach to reliability testing here. Yes, to calculate the signal-to-noise ratio (SNR) of the risk-adjusted IOH measure for each clinician, we adopted a multi-level hierarchical regression approach described by Morris (1983) to estimate the signal and noise. More information is provided in the response to Issue 12.

  Reference: Morris, C. N. "Parametric Empirical Bayes Inference: Theory and Applications." Journal of the American Statistical Association, vol. 78, no. 381, 1983, pp. 47–55.

- **Issue 4 (Q2 R4):** Specifications. More information is needed on how to deal with extreme blood pressure values from the anesthesia information management system (AIMS).

  - **Developer Response 4:** We are enclosing the full measure specifications, which provide guidance on how to deal with extreme/artifactual blood pressure values from the AIMS. The relevant excerpt from the Guidance section of the specification is as follows:

    "Because longitudinal blood pressure data can contain artifactual values (for example, inaccurate readings caused by the surgeon's leaning on the blood pressure cuff), the measure will drop MAP, SBP, and DBP readings that are likely to be artifacts. Specifically, the measure will drop individual MAP readings that meet any of the following criteria:
    - Documented as an artifact by the clinician
    - SBP ≥ 300 mmHg or ≤ 20 mmHg
    - DBP ≤ 5 mmHg or DBP ≥ 225 mmHg
    - SBP and DBP within 5 mmHg
    - MAP ≤ 30 mmHg or ≥ 250 mmHg"

- **Issue 5 (Q2 R4, Q11 R4):** Specifications. More information is needed on provider attribution and how to account for multiple providers on the same case.

  Developer Response 5: When multiple clinicians work on a case, either as a team or sequentially in a hand-off, the full case will be attributed to each of the clinicians. We are enclosing the full measure specifications, which provide guidance on provider attribution. The relevant excerpt from the Guidance section of the specification is as follows:

  "The measure attributes the full case to all reporting clinicians who provide care during any portion of the case from the beginning to the end of the measurement period."

- **Issue 6 (Q2 R4):** Specifications. More information is needed on the definition of and acquisition strategy for baseline MAP, one of the exclusion criteria.

    - **Developer Response 6:** We are enclosing the full measure specifications, which provide the measure's definition of "baseline MAP < 65 mmHg." The relevant excerpt from the Definitions section of the specification is as follows:

      "Baseline MAP < 65 mmHg: Cases in which the baseline MAP is below 65 mmHg. If one or more MAP values are available from the pre-operative holding area, the most recent value determines whether the patient meets the exclusion criteria. If no pre-operative holding area values are available, then the most recent pre-induction value from the operating room determines whether the patient meets the exclusion criteria. If a MAP reading is not available, then a calculated MAP value based on SBP and DBP readings is acceptable."[1]

      Our approach to determine a feasible way to define baseline MAP was informed by discussions with our clinical expert work group. We were advised that it is not consistently feasible to access blood pressure data from a primary care setting or from the preoperative H&P visit. The expert work group recommended using baseline blood pressure values from the pre-operative holding area, and if that's not available, the last reading in the operating room prior to induction, as both will be available to the anesthesiologist.

- **Issue 7 (Q2 R4):** Specifications. How does the measure deal with cases that require induced hypotension?

    - **Developer Response 7:** The measure does not currently exclude or adjust for cases that require induced hypotension. We are open to considering this as a denominator exclusion in future iterations of the measure, provided there is a low-burden and feasible way to identify induced hypotension in the anesthesia record.

- **Issue 8 (Q2 R5):** Specifications. Measurement period was not specified.

    - **Developer Response 8:** The measurement period is 12 months. We are enclosing the full measure specifications, which mention the measurement period, although we see now that is needs to be stated more prominently. The relevant reference in the "Steps for Calculating Unadjusted and Risk-Adjusted Measure Scores" section of the specification is as follows:

[1] From an earlier section of the specification: "If a clinician does not have MAP values available to report either for the baseline MAP or for measurements across the measurement period, the clinician may submit pairs of systolic and diastolic blood pressures (SBPs and DBPs) as a replacement for the MAP. The registry collecting the data will use these systolic and diastolic pressure values to calculate MAP values. Specifically, the registry will calculate MAP using the following formula: MAP = 1/3 (SBP) + 2/3 (DBP) (Sesso et al. 2000)."

"Run the measure on all anesthesia cases during the measurement period, representing a full calendar year."

- **Issue 9 (Q2 R6):** Specifications. Allowing the blood pressure reading to be taken using either invasive or non-invasive means introduces a source of ambiguity.

    - **Developer Response 9:** We are enclosing the full measure specifications, which provide guidance on how to calculate the measure in cases where both an invasive and non-invasive method is used. The relevant excerpts are as follows:

      Guidance section: "If the reporting clinician monitors a patient using more than one method and there are two MAPs available at the same point in time, the measure uses the invasive value for scoring the measure."

      Data elements and definitions section: "MAP < 65 mmHg: Refers to periods of time (minutes) in which the AIMS records a MAP reading that falls below 65 mmHg at any point between the anesthesia start time and anesthesia end time. The reading can be taken using either invasive or non-invasive means. If two readings are taken at the same time using a combination of invasive and non-invasive means, the invasive reading is used in calculating the measure..."

- **Issue 10 (Q2 R6):** Specifications. A given reading will only be carried forward for a maximum of five minutes. Clarify what happens if a patient has had an MAP < 65 for 14 minutes and there is no new reading at the critical time.

    - **Developer Response 10:** In the reviewer's example case where there is a reading of MAP < 65 then no reading for at least 14 minutes after, the initial reading would only count as 5 minutes of time toward the numerator threshold. If there were other instances of MAP < 65 during the case, and the cumulative total was for 15 minutes or more, the case would meet numerator criteria.

      The clinical recommendation is that blood pressure be taken at least every 5 minutes (ASA 2015). When invasive methods are used, it is common for BP readings to be recorded in AIMS every 1 to 3 minutes. In our experience working with testing data from 2 sites, we commonly saw BP readings every 1 to 5 minutes. Anesthesiologists who provided input to the measure development process also indicated that they may take more frequent readings if the blood pressure drops or is unstable. We are reassured that the example above, with long amounts of time between readings, is not expected to occur frequently.

      Reference: American Society of Anesthesiologists. (2015) "Standard for Basic Anesthetic Monitoring." Available at: https://www.asahq.org/standards-and-guidelines/standards-for-basic-anesthetic-monitoring

- **Issue 11 (Q2 R7):** Specifications. No measure specifications were released for review.

    - **Developer Response 11:** We are enclosing the full measure specifications document.

- **Issue 12 (Q6 R1, Q6 R7, Q8 R3, Q11 R1, Q11 R7):** Reliability testing. Methods of testing are unclear. Provide more information on the method and the formulas used for reliability testing.

    - **Developer Response 12:** To calculate the signal-to-noise ratio (SNR) of the risk-adjusted measure score (O:E ratio) for each clinician, we adopted a multi-level hierarchical regression approach to estimate the signal and noise separately. Specifically, we first estimated the "noise" (within-clinician variability) by calculating the variance of the $\frac{\sum O_i}{\sum E_i} \cdot \bar{Y}$ within each clinician, where the randomness is contributed by the observed event (i.e., MAP below 65 mmHg for cumulative total of 15 minutes or more) of each case within the clinician. Under the logistic regression setting of the risk adjustment model, the noise can be represented as $\bar{Y}^2 \times \frac{\sum_{i=1}^{n_k} Var(Y_k^i)}{\left(\sum_{i=1}^{n_k} E(Y_k^i)\right)^2} = \bar{Y}^2 \times \frac{\sum_{i=1}^{n_k} p_i(1-p_i)}{\left(\sum_{i=1}^{n_k} p_i\right)^2}$, where p_i is the estimated probability of surgery i with an event and n_k is the number of surgeries for clinician k.

      We next estimated the "signal" (between-clinician variance) iteratively, using a maximum likelihood estimation approach described by Morris (1983). This approach is appropriate for continuous measures such as an O:E ratio, and is analogous to the beta-binomial method used for binary outcomes (i.e., proportion measures). We computed the SNR statistic, *R*, as the ratio of the signal variance (which is common across all entities) to the sum of the signal variance and the noise variance (which varies by entity): $R = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$

      Reference: Morris, C. N. "Parametric Empirical Bayes Inference: Theory and Applications." Journal of the American Statistical Association, vol. 78, no. 381, 1983, pp. 47–55.

- **Issue 13 (Q6 R1, Q11 R1):** Reliability testing. Clarify whether risk-adjusted or unadjusted scores were used in reliability testing.

    - **Developer Response 13:** Reliability testing was conducted on the risk-adjusted, clinician-level scores (O:E ratios), since that is how the measure will be reported.

- **Issue 14 (Q6 R3, Q7 R2, Q7 R3, Q8 R3, Q11 R5, Q29 R2):** Reliability testing. Provide reliability estimates by provider volume, particularly for low volume providers. Also show distribution of reliability below the 25th percentile.

- **Developer Response 14:** Below are new reliability tables showing the distribution of reliability coefficients for the risk-adjusted measure (O:E ratio), by provider volume, in 2016 and 2017. As expected, reliability is directly associated with denominator size. In both years, around 75 percent of the clinicians had at least 101 cases in their denominator; this subgroup had high reliability of the IOH measure. Only 12 to 13 percent of clinicians (depending on the year) had fewer than 30 cases; reliability was below an acceptable threshold in this small subsample, as expected.

Distribution of reliability coefficients for risk-adjusted IOH measure in 2016, by number of denominator cases

	Clinicians with 1-30 cases	**Clinicians with 31-100 cases**	**Clinicians with 101-200 cases**	**Clinicians with 201-500 cases**	**Clinicians with 501+ cases**
Number of clinicians	78	87	78	216	207
Percent of sample (N = 666 clinicians)	12%	13%	12%	32%	31%
Mean	0.50	0.83	0.94	0.97	0.98
5th ptile	0.04	0.71	0.90	0.95	0.97
10th	0.11	0.73	0.91	0.95	0.98
20th	0.20	0.77	0.93	0.96	0.98
30th	0.26	0.79	0.93	0.96	0.98
40th	0.37	0.82	0.93	0.97	0.98
50th	0.53	0.84	0.94	0.97	0.99
60th	0.58	0.86	0.95	0.97	0.99
70th	0.66	0.87	0.95	0.98	0.99
80th	0.71	0.89	0.97	0.98	0.99
90th	0.74	0.91	0.97	0.99	0.99

Distribution of reliability coefficients for risk-adjusted IOH measure in 2017, by number of denominator cases

	Clinicians with 1-30 cases	**Clinicians with 31-100 cases**	**Clinicians with 101-200 cases**	**Clinicians with 201-500 cases**	**Clinicians with 501+ cases**
Number of clinicians	88	85	102	208	215
Percent of sample (N = 698 clinicians)	13%	12%	15%	30%	31%
Mean	0.39	0.81	0.92	0.96	0.98
5th ptile	0.07	0.67	0.85	0.93	0.97
10th	0.09	0.69	0.88	0.94	0.97
20th	0.14	0.73	0.90	0.95	0.97
30th	0.25	0.79	0.91	0.96	0.98
40th	0.36	0.81	0.91	0.96	0.98
50th	0.43	0.83	0.92	0.96	0.98
60th	0.48	0.84	0.93	0.97	0.98
70th	0.54	0.85	0.93	0.97	0.99
80th	0.61	0.87	0.96	0.98	0.99
90th	0.66	0.89	0.97	0.98	0.99

- **Issue 15 (Q7 R1):** Reliability testing. Add descriptive statistics on the number of cases per clinician.

    - **Developer Response 15:** The new tables below shows the distribution of the number of cases per clinician in 2016 and 2017. The median clinician-level denominator size is 294 cases in 2016, and 296 cases in 2017.

Distribution of clinician-level denominator size in 2016 (N = 666)

	Denominator size
Mean (sd)	357 (355)
5th ptile	9
10th	26
20th	65
30th	144
40th	221
50th	294
60th	396
70th	518
80th	626
90th	740
95th	796

Distribution of clinician-level denominator size in 2017 (N = 698)

	Denominator size
Mean (sd)	365 (367)
5th ptile	9
10th	23
20th	73
30th	138
40th	206
50th	296
60th	406
70th	507
80th	650
90th	794
95th	845

- **Issue 16 (Q11 R2):** Reliability testing. Consider also using a secondary method to assess reliability, such as ICC. Median reliability metrics in excess of 0.90 for denominators less than 100 seem implausibly high.

    - **Developer Response 16:** The original reliability tables did not present the reliability coefficients separately for providers with denominators less than 100; they showed reliability coefficients for providers that met various minimum denominator thresholds. The new reliability tables attached to Issue 14 show reliability coefficients by case volume. The median reliability coefficient was 0.35 to 0.37 (depending on the year) for providers with a denominator of 1 to 30 cases

and 0.68 to 0.75 for providers with a denominator of 31 to 100 cases. We have not yet been able to conduct reliability testing using a secondary method, but we hope that the new reliability tables are granular enough to give the reviewers confidence in the results.

Validity

- **Issue 1 (Q12 R1):** Exclusions. The high rate of missing data for the baseline MAP exclusion in Site 1 raises concerns about the validity of this exclusion and results of exclusion testing. Retest this exclusion when more data are available.

    - **Developer Response 1:** We agree with the reviewer's suggestion to retest this exclusion in the future when more data are available.

- **Issue 2 (Q12 R5):** Exclusions. The prevalence of the exclusion criteria is low, and the exclusions do not change the overall measure score. More information is required about whether individual clinician's scores change as a result of exclusions.

    - **Developer Response 2:** In a new analysis, we calculated the clinician-level risk-adjusted scores (O:E ratios) with and without the denominator exclusions. We looked at the difference between scores for each clinician. The new table below shows the distribution of the difference between the clinician's risk-adjusted scores with versus without exclusions.

      The table shows that, on average, the exclusions had no impact on clinicians' risk-adjusted scores. However, the maximum difference of 1.49 suggests that there are at least some clinicians for whom the exclusions made a meaningful difference in their score.

Distribution of the difference in clinicians' risk-adjusted scores (O:E ratios), with versus without denominator exclusions (N = 833)

	Difference in O:E ratio
Mean	0.00
Minimum	-0.27
5th ptile	-0.06
10th	-0.04
20th	-0.02
30th	0.00
40th	0.00
50th	0.00
60th	0.00
70th	0.01
80th	0.01

	Difference in O:E ratio
90th	0.03
95th	0.04
Maximum	1.49

- **Issue 3 (Q12 R5, Q26 R5):** Exclusions. Provide further rationale for each exclusion.

    - **Developer Response 3:** Our approach to selecting denominator exclusions was informed by a review of the literature and discussions with our clinical expert work group. The measure excludes cases on the extreme ends of risk for IOH. Liver and lung transplants are both excluded because they are highly complex surgeries that involve interruption of the circulatory system, leading to blood pressure variation unrelated to the actions of the anesthesiologist. Cataract surgeries are excluded for the opposite reason: they are short and fairly non-invasive, with very little risk of IOH.

      Obstetric non-operative procedures are excluded because they are non-operative and because blood pressure is naturally lower in pregnancy. The evidence base on MAP thresholds for IOH are not necessarily generalizable to pregnant patients.

      Cases with ASA status classification V or VI are excluded because the patients are either not expected to survive without surgery (ASA V) or because they are brain dead (ASA VI), making IOH a lower priority concern.

      Cases in which the patient's baseline MAP is less than 65 prior to induction are excluded. The purpose of the exclusion is to remove cases who are entering the anesthesiologist's care already in a state that is counting toward the numerator.

- **Issue 4 (Q13 R1):** Meaningful differences in performance. Please add n's to Tables 16 and 17.

    - **Developer Response 4:** Tables 16 and 17 in the testing form show the distribution of clinician-level unadjusted and risk-adjusted scores, and the percent of clinicians with risk-adjusted scores (O:E ratios) that are statistically significantly different from 1.0. Both tables are based on all testing data combined (n = 178,343 cases after exclusions). They include measure scores from 833 unique clinicians across multiple years, for a total of 1,631 clinician-level scores represented.

- **Issue 5 (Q13 R7):** Meaningful difference in performance. A large proportion of clinicians had O:E ratios that were statistically significantly different than 1.0. Describe the method for calculating the confidence interval around the clinician-level score.

    - **Developer Response 5:** We estimated the confidence interval of the O:E ratio in two steps. First, under the logistic regression setting of the risk adjustment model, the standard error of clinician k can be represented as $SE_k = Var\left(\frac{O_k}{E_k}\right) = \frac{\sum_{i=1}^{n_k} Var(Y_k^i)}{\left(\sum_{i=1}^{n_k} E(Y_k^i)\right)^2} = \frac{\sum_{i=1}^{n_k} p_i(1-p_i)}{\left(\sum_{i=1}^{n_k} p_i\right)^2}$, where p_i is the estimated probability of surgery i with an event and n_k is the number of surgeries for clinician k. Second, the 95% confidence interval of clinician k is $(\frac{O_k}{E_k} - 1.96 * SE_k, \frac{O_k}{E_k} + 1.96 * SE_k)$.

- **Issue 6 (Q14 R1, Q14 R2, Q14 R7):** Comparability of performance scores when more than one set of specifications. This section of the testing form is not applicable to the measure, and the analysis included (correlation of adjusted and unadjusted scores) does not seem to belong in this section.

    - **Developer Response 6:** Apologies for the confusion. The measure does not have multiple specifications or use multiple data sources, so we may not have needed to complete this section. The testing form instructions note that the section is relevant to measures that are risk adjusted or measures with more than one set of specifications/instructions. We included a correlation plot to show the association between each clinician's unadjusted and risk-adjusted score, in case it was perceived that the steps for calculating the unadjusted and adjusted scores count as two sets of instructions. Thank you for clarifying that it is not needed.

- **Issue 7 (Q14 R4):** Comparability of performance scores when more than one set of specifications. Discuss how different methods of blood pressure measurement (for example, invasive arterial line versus non-invasive blood pressure cuff) may affect the measure score itself.

    - **Developer Response 7:** The blood pressure data for the measure comes from the AIMS system regardless of whether invasive or non-invasive readings are taken. In that sense, the measure uses only one data source. We acknowledge that some clinical studies show that invasive and non-invasive readings may not always align; however, both methods are accepted as valid and clinically-appropriate assessments of patient blood pressure and are therefore allowable in the context of this measure. The measure uses invasive readings in cases in which both invasive and non-invasive readings are taken.

- **Issue 8 (Q15 R1):** Missing data. Table 19 in the testing form shows the IOH rate by missingness of BMI data. Clarify whether similar analyses were conducted for other variables and both sites.

    - **Developer Response 8:** BMI was the only risk adjustment variable with missing data in either site; therefore, we did not need to conduct similar analyses using other risk adjustment variables.

      We added a similar analysis related to missing data for the baseline MAP exclusion, which was missing from the majority of cases at Site 1. The new table below shows the rates of IOH among cases with missing versus non-missing data for low baseline MAP (the exclusion criterion) in Site 1. It shows that cases with missing data for that exclusion are less likely to go on to develop IOH than cases with non-missing data. Since this appears to be a low risk group, we are less concerned that 75 percent of cases had missing baseline MAP and therefore we could not assess the exclusion.

Rates of IOH among Site 1 cases based on availability of baseline MAP (exclusion criterion)

Availability of baseline MAP	Frequency	IOH rate
Available	43,053	33%
Missing	82,153	21%

- **Issue 9 (Q15 R1):** Missing data. Clarify why the counts in Tables 18 and 19 are larger than the overall counts for each site.

    - **Developer Response 9:** Table 18 in the testing form shows the availability of baseline blood pressure data among Site 1 anesthesia cases. The table shows 133,357 cases, which is greater than the 125,206 cases from Site 1 that met the initial patient population. The analyses of missing blood pressure variables were done in the raw vital signs datasets that the test sites sent us, prior to cleaning the data and restricting to the initial patient population.

      Table 19 in the testing form shows the rates of IOH among cases in Site 2 that have missing BMI values versus non-missing BMI values. The table was mistakenly shown at the clinician-case level (n=183,156), rather than the case-level. This means that cases with multiple clinicians on it were represented multiple times in the table. We have corrected the frequencies in table below to be at the case level, with each case represented once. The IOH rates and interpretation do not change: unadjusted rates of IOH are 3 percent higher among those with missing BMI.

Table 19: Rates of IOH among Site 2 cases based on availability of BMI data

Available of BMI	Frequency	IOH rate
Available	58,267	47.0%
Missing	2,446	49.9%

- **Issue 10 (Q15 R1, Q15 R7):** Missing data. Baseline MAP is missing from a large proportion of cases. The prevalence of the exclusion for baseline MAP < 65 (0.9%) among the cases with non-missing data may not be representative of the true prevalence of the exclusion criterion. Please clarify whether cases missing baseline MAP may be more likely to develop IOH than cases with non-missing baseline MAP.

    - **Developer Response 10:** We agree with the reviewer—we cannot determine whether the rate of exclusion for low baseline MAP for cases with a missing baseline MAP would be the same as the rate (0.9%) for cases with a non-missing baseline MAP. However, the new table included in Issue 8 above shows that cases with a missing baseline MAP go on to have a lower rate of IOH than cases with a non-missing baseline MAP, implying they are a lower risk population. Based on that, we do not expect that cases with a missing baseline MAP would have a higher rate of exclusion for low baseline MAP than 0.9%.

- **Issue 11 (Q15 R1):** Missing data. The 3% different in IOH rate between cases with missing and non-missing BMI (a risk adjustment variable) is not necessarily "slight" as classified in the testing form.

    - **Developer Response 11:** The 3% absolute difference in IOH rate between those with non-missing BMI values (47.0% IOH rate) and missing BMI values (49.9% IOH rate) seemed qualitatively small to us, given the very high prevalence of IOH in both groups and the large variation in IOH by clinician. The difference represents a 6% relative increase in IOH. We agree with the reviewer that the decision of whether a difference in IOH prevalence is clinically meaningful or not is subjective, and not everyone would agree that the difference between a 47% prevalence and a 50% prevalence is slight.

- **Issue 12 (Q15 R5, Q26 R5):** Missing data. Clarify how the measure will be applied to cases with missing risk-adjustment data (for example, BMI).

    - **Developer Response 12:** The risk-adjustment model that is used to calculate the expected number of cases of IOH per clinician uses case-wise deletion. It will drop any case that is missing one or more of the five risk-adjustment variables. The provider-level, risk-adjusted score will therefore be based on their subset of cases that have complete data on the risk factors.

      The low rates of missingness among the risk-adjustment variables, coupled with the large denominator size per clinician imply that case-wise deletion is unlikely to introduce bias into the risk-adjusted scores. This issue is one reason why data

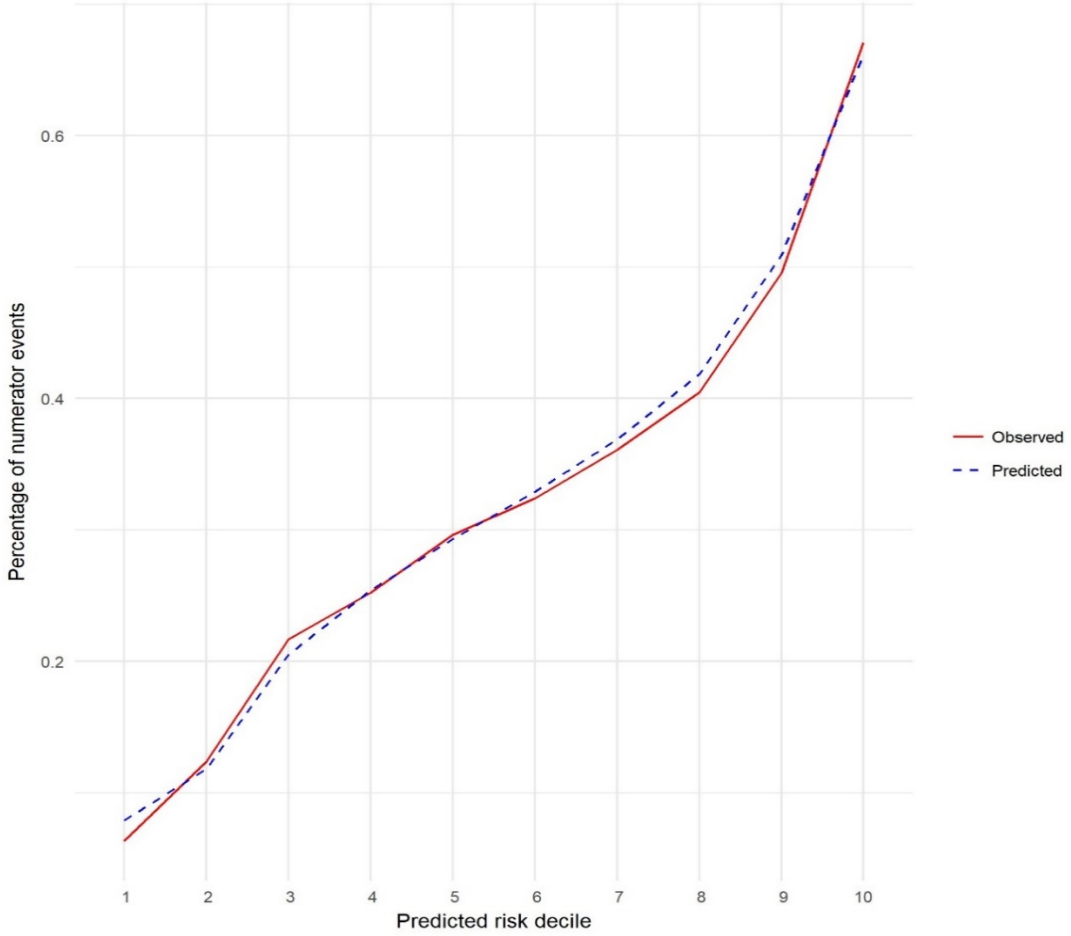
availability and feasibility were important criteria when we selected the evidence-based risk factors to include in the risk-adjustment model.

- **Issue 13 (Q16 R3, Q26 R3):** Risk adjustment. The model calibration plot is shown for the full dataset rather than just for the validation half-sample.

    - **Developer Response 13:** Apologies for this important oversight. As the reviewer suggests, we inadvertently included the calibration plot and decile table for the full sample rather than just the validation half-sample that was not used for model development. We ran them on both samples and included the less important output in our testing form. Thank you for the opportunity to provide the additional results.

      Below are the calibration plot and decline table among the validation half-sample that was not used to develop the model. The interpretation is the same as in the testing form: the model shows good calibration at all levels of risk.

Observed and predicted IOH by risk decile, validation sample (n = 89,171 cases)



Observed and predicted numerator percentages by risk decile, validation sample (n = 89,171 cases)

Decile	Count of Anesthesia Cases	Mean Observed IOH Rate	Mean Predicted IOH Rate
1st	8,918	6.3%	7.9%
2nd	8,917	12.4%	11.8%
3rd	8,917	21.7%	20.5%
4th	8,917	25.2%	25.4%
5th	8,917	29.6%	29.3%
6th	8,917	32.4%	32.9%
7th	8,917	36.1%	36.9%
8th	8,917	40.5%	41.9%
9th	8,917	49.6%	50.9%
10th	8,917	67.1%	66.2%

- **Issue 14 (Q16 R1):** Risk adjustment. Was low baseline MAP considered as a risk factor rather than a denominator exclusion?

    - **Developer Response 14:** We did not consider using low baseline MAP (< 65) as a risk factor rather than a denominator exclusion. The purpose of the exclusion is to remove cases who are entering the anesthesiologist's care already in a state that is counting toward the numerator. It functions similar to an exclusion for a condition 'present on arrival.' Based on that rationale, we always planned for it to be an exclusion rather than a risk factor. Our expert work group agreed with the rationale for the exclusion.

- **Issue 15 (Q16 R1):** Risk adjustment. Provide descriptive statistics for the five risk adjustment variables.

    - **Developer Response 15:** The new table below shows the descriptive statistics for the five risk factors used in the risk adjustment model. During model development we used data visualizations and model fit statistics to determine which form of the variables (e.g. continuous or categorical) were most appropriate to reflect the relationship between the predictor and the outcome (IOH). The final model includes age and BMI as continuous variables, sex as binary, and ASA physical classification status and surgery length as categorical.

Descriptive statistics on the risk factors, in the development/training sample (n = 89,172 cases)

Risk factor			
	Mean (sd)	**Range (min - max)**	**Interquartile range**
Age (years)	56.9 (16.8)	18 - 108	24.2
BMI	29.6 (7.8)	10.8 – 182[a]	9.1
	n (Percent)		
ASA status			

Risk factor			
	Mean (sd)	**Range (min - max)**	**Interquartile range**
ASA I	3,144 (3.5%)		
ASA II	27,668 (31.0%)		
ASA III (reference)	45,143 (50.6%)		
ASA IV	13,217 (14.8%)		
Female	47,877 (53.7%)		
Surgery length			
< 60 minutes (reference)	18,807 (21.1%)		
60-119 minutes	26,986 (30.3%)		
120-179 minutes	16,702 (18.7%)		
180-239 minutes	9,913 (11.1%)		
240-299 minutes	6,377 (7.2%)		
300+ minutes	10,387 (11.6%)		

[a]BMI values of less than 10 and greater than 200 are considered implausible and are removed prior to measure calculation. In the full dataset, 37 such cases were removed prior to calculating or testing the measure.

- **Issue 16 (Q16 R1):** Risk adjustment. Add the test statistics, p-values, and 95% confidence intervals to Table 11, which shows the regression coefficients for the risk adjustment model. If not statistically significant, provide further justification for including these variables in the model.

    - **Developer Response 16:** The new table below adds columns to the original Table 11, to show the 95% confidence interval (CI), test statistic, and p value for each model coefficient. All coefficients are statistically significant.

Risk-adjustment model coefficients, in the model development/training sample (n = 89,172 cases)

Parameter	Value	95% CI	Statistic	p value
β_0: Constant/Intercept	-1.482	-1.583, -1.380	-28.528	0.000
β_1: Coefficient 1: Age	-0.008	-0.009, -0.007	-15.976	0.000
β_2: Coefficient 2: ASA_1	0.400	0.316, 0.484	9.321	0.000
β_3: Coefficient 3: ASA_2	0.164	0.128, 0.201	8.777	0.000
β_4: Coefficient 4: ASA_4	0.532	0.488, 0.576	23.676	0.000
β_5: Coefficient 5: BMI	-0.018	-0.020, -0.016	-17.741	0.000
β_6: Coefficient 6: Surg_Length_Cat_60–119	1.231	1.175, 1.286	43.754	0.000
β_7: Coefficient 7: Surg_Length_Cat_120–179	1.664	1.606, 1.722	56.479	0.000
β_8: Coefficient 8: Surg_Length_Cat_180–239	1.871	1.808, 1.934	58.189	0.000
β_9: Coefficient 9: Surg_Length_Cat_240–299	2.128	2.059, 2.198	60.055	0.000

Parameter	Value	95% CI	Statistic	p value
β_{10}: Coefficient 10: Surg_Length_Cat_300–	2.810	2.746, 2.874	86.526	0.000
β_{11}: Coefficient 11: Female1	0.171	0.141, 0.202	10.914	0.000

- **Issue 17 (Q16 R1):** Risk adjustment. Cases with ASA I are at increased risk of having IOH. Might this association change if the model was adjusted for resting blood pressure?

    - **Developer Response 17:** Yes, it is plausible that the association between IOH and ASA I could change if the model adjusted for resting blood pressure. Unfortunately we are not able to test that in the model due to lack of data. There is no agreed upon definition for resting blood pressure, and blood pressure values from prior to the day of surgery are not reliably available in the anesthesia record. In a small number of testing cases, we had access to blood pressure values from the preoperative H&P visit. Regular access to that data would likely require linking to the EHR of the patient's primary care provider or other ambulatory care provider, which introduces feasibility challenges and burden.

- **Issue 18 (Q16 R1):** Risk adjustment. Provide values for each risk adjustment variable in Table 12.

    - **Developer Response 18:** Table 12 shows the associations between each risk adjustment variable and IOH. The risk adjustment variables were each tested using the same functional form and categories as used in the risk adjustment model. Age and BMI were tested as continuous variables and sex was binary. ASA physical status classification was tested as four categories: ASA I, ASA II, ASA III, and ASA IV. Surgery length was tested as six categories: < 60 minutes, 60-119 minutes, 120-179 minutes, 180-239 minutes, 240-299 minutes, and >= 300 minutes.

- **Issue 19 (Q16 R1):** Risk adjustment. Provide results that show the O:E ratio in the model validation sample, by subgroup (age, sex, BMI, ASA status, surgery length), to demonstrate that the model was equally predictive for different patient populations.

    - **Developer Response 19:** The table below shows the O:E ratios for different patient subgroups. These analyses were conducted during model development and validation but were not presented in the original NQF testing form. The O:E ratios range from 0.93 to 1.04, which provides evidence that the model is well calibrated in different patient populations.

O:E ratios by patient subgroup, validation sample (n = 89,171 cases)

Subgroup	O:E ratio
By age category	
18-39 years	1.02

Subgroup	O:E ratio
40-64 years	0.95
65-74 years	1.02
75+ years	1.04
By BMI	
Underweight and normal (BMI <25)	1.03
Overweight and obese (BMI >= 25)	0.98
By ASA status	
ASA I	0.97
ASA II	0.99
ASA III	0.99
ASA IV	0.99
By Sex	
Female	0.99
Male	0.99
By surgery length	
<60 minutes	0.93
60-120 minutes	0.98
>120 minutes	1.00

- **Issue 20 (Q16 R1):** Risk adjustment. It would be helpful to have a clear description of the method used to compute the clinician-level risk-adjusted score included in the testing form, acknowledging that it is clearly described in the measure specifications.

    - **Developer Response 20:** Please see the enclosed specification document, pages 6-12, for step-by-step instructions for calculating the clinician-level risk-adjusted score.

- **Issue 21 (Q11 R4; Q13 R4; Q16 R4, Q26 R4):** Risk adjustment. The risk adjustment model may have inadequate case mix adjustment, for example if anesthesia providers work on different types of surgical cases. The analysis does not demonstrate how the effect of the anesthesia provider can be isolated from the effect of surgical procedure, patient factors, or surgeon.

    - **Developer Response 21:** The measure is currently adjusted for surgery length as a proxy for the effect of the complexity of the surgical procedure, and it is adjusted for several patient factors: ASA physical classification status (as a proxy of the degree of systemic disease), age, BMI, and sex. We are open to considering other risk factors in future iterations of the risk adjustment model, provided there is a low-burden and feasible way to collect the information for

all/nearly all cases. We would consider risk adjusting for the surgical procedure type, using a standard classification system, in future updates to the measure.

- **Issue 22 (Q16 R5, Q16 R7, Q26 R5):** Risk adjustment. One of the risk adjustment variables—length of surgery—is not available before the surgery. It's possible the direction of causality could be bidirectional or reversed (e.g., having IOH could lead to a longer surgery). There may be other ways to proxy for complexity of the surgery and exposure time using a variable available prior to surgery (e.g., average length of surgery for that specific procedure, surgery type, or RVU).

    - **Developer Response 22:** There is a strong association between surgery length and IOH, which we believe is largely driven by longer surgeries being more complex and having more exposure time under anesthesia. We agree that an episode of IOH could possibly lengthen the surgery, for example if extra interventions are required. However, in discussions with clinicians, we were advised that in most cases, episodes of IOH will likely not delay or stop the surgery unless the IOH is life threatening.

      To minimize any potential effect of bidirectional association, we have modeled surgery length categorically in the risk adjustment model, with each category representing 60 minutes, which makes it less likely that an episode of IOH would cause a case to move from one surgery length category to another. Therefore, even if IOH lengthens the surgery slightly, rarely would it affect the clinician's expected number of IOH cases that is used to calculate his O:E ratio.

      As mentioned earlier, we are open to considering other risk factors in future iterations of the risk adjustment model, provided there is a low-burden and feasible way to collect the information for all/nearly all cases. We would consider risk adjusting for the surgical procedure type, using a standard classification system, in future updates to the measure.

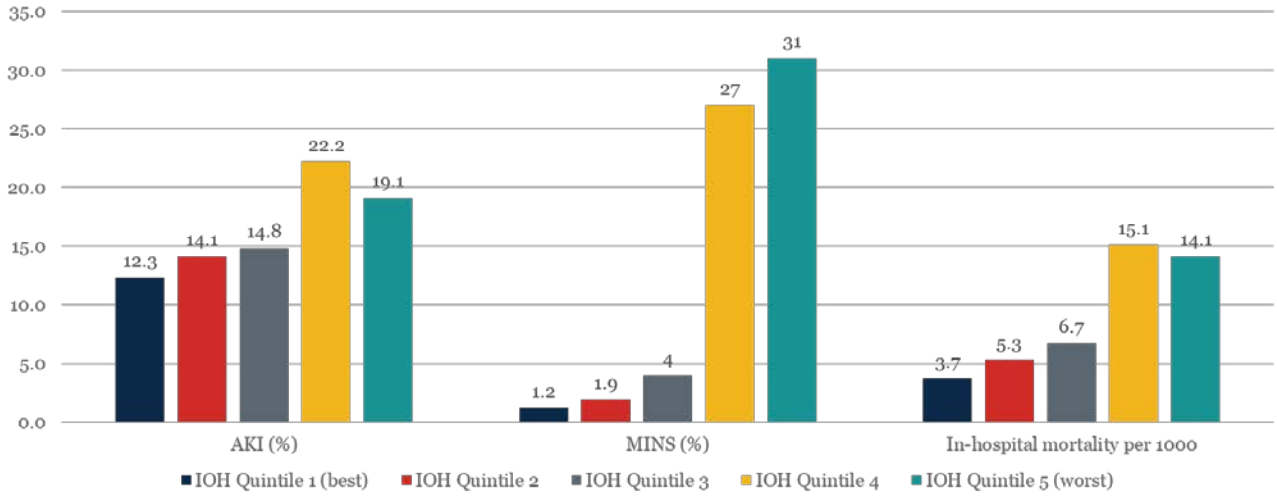
- **Issue 23 (Q21 R1, Q21 R3, Q21 R5, Q26 R1, Q26 R7):** Validity testing. The validity analysis was conducted at the case level (associations between the case-level numerator and each adverse outcome). It would be stronger to conduct validity testing using the risk-adjusted, clinician-level score.

    - **Developer Response 23:** The new graph below provides evidence of the validity of the risk-adjusted, clinician-level IOH score. We calculated the rates of adverse patient outcomes (acute kidney injury [AKI], myocardial injury after noncardiac surgery [MINS], in-hospital mortality), by risk-adjusted IOH score quintile. This analysis was done in Site 1 only, as Site 2 did not provide data on AKI or MINS.

      The graph shows that clinicians with the worst 40 percent of risk-adjusted scores (quintiles 4–5) have meaningfully higher rates of AKI, MINS, and in-hospital

mortality compared with clinicians with the best 60 percent of scores (quintiles 1–3). The relationship is particularly strong for MINS and in-hospital mortality.

Incidence of AKI, MINS, and in-hospital mortality, by clinician-level, risk-adjusted IOH score quintile



- **Issue 24 (Q22 R1, Q26 R1):** Validity testing. Clarify the sample sizes in Table 7 (predictive validity) and Table 8 (known group validity). They appear to be subsamples, but section 1.7 does not describe different testing subsamples.

    - **Developer Response 24:** Tables 7 and 8 show the IOH incidence in different subgroups (e.g., those with and without AKI; those under age 65 versus those 65 and older). They were conducted on the full sample (n = 178,343 cases after exclusions). The exception is noted in the footnote of Table 7 that the AKI and MINS analyses were only conducted in Site 1, as Site 2 did not provide those variables. Apologies for not also noting this analytic subsample in section 1.7.

      These tables may have looked like subsamples because we show the IOH incidence numerator counts in the column “IOH count,” but we do not show the denominator counts (that is, the number of with and without AKI, among whom we assessed IOH incidence) that could be used to confirm sample size.