NATIONAL QUALITY FORUM

Fall 2018 Scientific Methods Panel Evaluation Summaries

Measure Evaluation Criteria Rating Key: H – High; M – Medium; L – Low; I – Insufficient

Subgroup 1: Initial Evaluation Call, Tuesday, October 9 from 2-4 pm ET; Follow-Up Evaluation Call, Monday, October 22 from 10 am-12 pm ET

During its initial call, Subgroup 1 discussed four measures (0753, 2456, 1716, and 1717) and accepted the preliminary analysis decisions for two measures (0729 and 3474) without further discussion. One measure (3450) was pulled for discussion during the call and was discussed during the Subgroup's follow-up call. The final results for the seven measures evaluated by Subgroup 1 are presented below.

Measures Discussed by the Subgroup

0753 American College of Surgeons – Centers for Disease Control and Prevention (ACS-CDC) Harmonized Procedure Specific Surgical Site Infection (SSI) Outcome Measure

Scientific Methods Panel Votes: Consensus Not Reached on Reliability

- Reliability: H-0, M-2, L-1, I-2
- Validity: H-1, M-4, L-0, I-0

Reliability

- Testing was conducted at the data element and measure score levels.
- Limited data element **validity** testing was conducted for the population (state) level of analysis. If accepted as adequate data element validation, no additional reliability testing for the population (state) level of analysis is required**.**
- Measure Score (for facility level of analysis)
    - The developer stated that "*Reliability was estimated as the between- facility variance from a generalized linear mixed model divided by the total variance estimated from the same model*." The Methods Panel would have liked more detail about the methodology used in testing.
        - Results for colorectal surgeries were incomplete (which disturbed panel members) and mean reliability estimates were fairly low:
            - For COLO SSI, mean reliability=50.1%; xxx of 2,009 facilities met the Minimum Precision Criteria (MPC) and had reliability exceeding 40%. Developer noted that "Around one-third of facilities that met the MPC had reliability below the commonly-used 40% threshold for COLO SSI".
            - For HYST SSI, mean reliability=52.9%; and 652 of 787 facilities meeting the MPC had reliability exceeding 40%.
- Although not required by NQF at this time, Methods Panel members also noted that an analysis of the "stability" of the measure results for the facility level of analysis (i.e., that

facility scores and rankings would not change dramatically in the short term) would have been helpful due to the rarity of the outcomes being measured.

Validity

- This measure is risk adjusted and a limited amount of data element validity testing was conducted. Although the panel rated validity as "Moderate", members had several concerns.
- Methods Panel members voiced concern about exclusions to the denominator that are due to "data quality, data outliers, and data errors". Members were concerned with the assertion that "*missing data is not a problem*" if incomplete reporting by facilities (which may be related to quality) means that they are not being assessed by the measure.
- Many facilities did not meet the Minimum Precision Criteria (MPC), which requires a facility to have at least one predicted event from the risk-adjustment model. This underscores the rarity of the outcomes being assessed.
- Risk model calibration (2b3.7) for SSI HYST shows a potential issue with Hosmer Lemeshow test (p = 0.012). Panel members noted that decile plots could help illuminate whether there is a problem with model calibration.
- Data element testing (for facility and population level of analysis)
    - Conducted limited data element validity testing for 7 states for the COLO SSI outcome and for 3 states for the HYST SSI outcome. It was not completely clear what element(s) were tested (likely the numerator only).
    - Panel members also noted that the relatively low sensitivity results, noting that the submission indicates there are disincentives for reporting colon surgery SSIs to the NHSN
        - Results: COLO – Mean measurements identified (ranges) were as follows:
            - Sensitivity: 74.9% (59.8-90.1)
            - Specificity: 99.1 % (98.7-100)
            - Positive Predictive Value: 95.8% (91.7-100)
            - Negative Predictive Value: 93.5% (85.3-97.2)
        - Results: HYST – Mean measurements identified (ranges) were as follows:
            - Sensitivity: 80.7% (75.4-100)
            - Specificity: 98.9% (88.9-99.1)
            - Positive Predictive Value: 92.6% (91.5-94.4)
            - Negative Predictive Value: 96.9% (96.9-100)
- Developers did not conduct data element validation for the variables used in the risk-adjustment model. Panel members think these are critical data elements that should be tested.
- At least one panel member thought that the risk-adjustment approach might not be robust enough, but believes the Standing Committee should weigh in from a clinical perspective.
- Another panel member expressed frustration about lack of reporting of the "optimism statistic", which apparently would provide information about the utility of the risk-adjustment approach (the submission indicates this analysis was conducted but the results were not provided).

2456 Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient

Scientific Methods Panel Votes: Measure does not pass on Validity

- Reliability: H-0, M-4, L-1, I-0
- Validity: H-0, M-0, L-4, I-1

This measure was reviewed by NQF's Scientific Methods Panel. The panel agreed that the measure, as submitted, does not meet NQF's requirements for validity due to lack of empirical testing as well as concerns about the generalizability of the measure and the fairness of comparisons of measure results across facilities.

1716 National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Methicillin-resistant Staphylococcus aureus (MRSA) Bacteremia Outcome Measure

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0; M-5; L-0; I-0
- Validity: H-0; M-4; L-1; I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- The developer provided results of data element validity testing; NQF' guidance states that additional reliability testing is not needed if empirical validity testing of patient-level data is conducted and the results are adequate.
- Panel members noted that while the testing information met NQF's minimum requirements, they would have liked to see separate reliability testing of data elements.

Validity

- Validity testing was performed at the data element level.
- Data Element
    - The developers provided a summary of validation studies conducted in 5 states.
    - These studies involved taking a sample of charts from a sample of facilities in varying years; these samples were then reviewed by trained chart abstractors and compared against data reported to the National Healthcare Safety Network (NHSN).
    - Developer provides sensitivity/specificity/PPV/NPV seemingly for the MRSA variable only. Panel members also noted that testing of variables included in the risk adjustment model was not reported; no information is provided on the validity of data elements used for risk adjustment and to identify the denominator population.
    - Results:

		Sensitivity	Specificity	Positive predictive value	Negative predictive value
Tennessee	2015	80.9%	87.5%	97.5%	42.8%
Wisconsin	2009	95.2%	63.6%	93.7%	70%

		Sensitivity	Specificity	Positive predictive value	Negative predictive value
New Mexico	2016	98.7%	100%	100%	98.8%
California (MRSA/VRE BSI)	2014	88%	NP*	NP*	NP*
Maine	2015	83%	NP*	74%	NP*

*NP- Not provided

- Risk Adjustment
  - This is a risk-adjusted model with 6 risk factors: inpatient community onset prevalence; average length of stay, medical school affiliation; facility type; number of ICU beds; and outpatient community onset prevalence.
  - No social risk factors were included because these are not collected in the NHSN for all patients in the patient population.
  - The risk model was developed using negative binomial regression, in which risk factors were evaluated by both univariate and multivariate modeling steps. The multivariate regression model was confirmed and validated using bootstrap validation techniques.
  - Results:
    - The p-values for all variables in the final multivariate model were statistically significant, with several variables having a p-value < 0.0001.

1717 National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Clostridium difficile Infection (CDI) Outcome Measure

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0; M-5; L-0; I-0
- Validity: H-0; M-5; L-0; I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- The developer provided results of data element validity testing; NQF' guidance states that additional reliability testing is not needed if empirical validity testing of patient-level data is conducted and the results are adequate.
- Panel members noted that while the testing information met NQF's minimum requirements, they would have liked to see separate reliability testing of data elements.

Validity

- Validity testing was performed at the data element level.
- Data Element
  - The developers provided a summary of validation studies conducted in 5 states.
  - These studies involved taking a sample of charts from a sample of facilities in varying years; these samples were then reviewed by trained chart abstractors and compared against data reported to the National Healthcare Safety Network (NHSN).

- Developer provides sensitivity/specificity/PPV/NPV seemingly for the Clostridium difficile (C. diff) variable only. Panel members also noted that testing of variables included in the risk adjustment model was not reported; no information is provided on the validity of data elements used for risk adjustment and to identify the denominator population.
- Results:

State	Year of data validated	Records reviewed	sensitivity	specificity	PPV	NPV
Connecticut	2013	1085	93	99	99.9	75
Colorado	2015	359	95	100	100	80
Tennessee	2015	534	89.4	73.5	98	32
Utah	2016	394	92.5	100	100	42.9
New Mexico	2016	302	100	58.3	98.3	100
New York	2014	1787	89.4	100	100	42.8
Overall		**4461**	**94.9**	**93.7**	**99.4**	**58.7**

- Risk Adjustment
  - This is a risk-adjusted model with 7 risk factors: Inpatient community onset prevalence; CDI test type, medical school affiliation; number of ICU beds; facility type; facility bed size; and reporting from ED or 24-hour observation unit.
  - No social risk factors were included because these are not collected in the NHSN for all patients in the patient population.
  - The risk model was developed using negative binomial regression, in which risk factors were evaluated by both univariate and multivariate modeling steps. The multivariate regression model was confirmed and validated using bootstrap validation techniques.
  - Results:
    - The p-values for all variables in the final multivariate model were statistically significant, with several variables having a p-value < 0.0001.

3450 Practice Environment Scale - Nursing Work Index (PES-NWI) (composite and five subscales)

Scientific Methods Panel Votes: Measure passes

- Reliability: H-3, M-1, L-0, I-1
- Validity: H-3, M-1, L-0, I-1

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- Reliability was conducted at the data element and measure score level.
- Data element
    - Conducted by computing Cronbach's alpha.
    - Results: Provided overall summary of results based on 46 articles reviewed by Swiger et al (2017).
        - 37 articles reported Cronbach's alphas; coefficients ranged from .71 – .96, with the exception of one .67, and one .53 in a small sample size.
- Measure score
    - Conducted by assessing inter-rater reliability, which focuses on whether nurses give consistent responses within a hospital or nursing unit, as compared to across hospitals or nursing units in a sample. Performance measure score reliability is assessed using the intraclass correlation (ICC) (1,*k*),
    - Results: based on 14 articles below and the 2015 National Database of Nursing Quality Indicators nurse survey data

Reference	# organizational units (hospitals or nursing units)	# nurses	ICC (1,k) statistics reported or summarized
Lake (2002)	16 magnet hospitals proportionate by regions of the country	1,610	.88 to .97
Lake et al (2006)	156 adult community hospitals in Pennsylvania	10,962	.67 to .82
Clarke (2007)	188 Pennsylvania general acute care hospitals	11,512	.70 to .90
Flynn et al (2010)	63 Medicare and Medicaid certified nursing homes in New Jersey	897	Composite: .68 Subscales range: .55 to .75
Brooks-Carthon et al (2011)	429 hospitals across four states (Florida, Pennsylvania, New Jersey and California)	98,000	Subscales range: .73 to .90
McHugh et al (2012)	396 adult, non-federal acute care hospitals across four states (CA, FL, NJ, PA)	16,241	.61
Kelly et al (2013)	320 hospitals across four states (CA, FL, NJ, PA)	3,217	.69
McHugh et al (2013)	564 Magnet and non-Magnet hospitals across four states (CA, FL, NJ, PA)	100,000	.81
Kelly et al (2014)	303 adult care hospitals across four states (CA, FL, NJ, PA)	55,159	.71
McHugh et al (2014)	534 hospitals across four states (CA, FL, NJ, PA)	26,005	.85
Carthon et al (2015)	419 acute care hospitals across three states (CA, FL, NJ, PA)	20,605	.74 to .91
Ma et al (2015)	373 hospitals from 44 states	33,845	Ranged from .80 to .87
Lake et al (2016)	171 hospitals across four states (CA, FL, NJ, PA)	1,247	4 subscales >.60; 5th = .58
Swiger et al (2018)	45 acute care units in 10 Army hospitals	180	ICC (1,*k*) reported as satisfactory

Analysis for 2018 NQF measure maintenance using 2015 National Database of Nursing Quality Indicators nurse survey data

Measure	ICC (1,k)
Subscale	
Collegial Nurse-Physician Relations	.936
Nursing Foundations for Quality Care	.966
Nurse Manager Ability, Leadership, and Support	.949
Nurse Participation in Hospital Affairs	.973
Staffing and Resource Adequacy	.967
Composite	.966

Note. N = 451 hospitals and from 157,481 to 157,522 staff nurses. ICC (1,k) estimated in one-way ANOVA.

- One panel member had concerns that the reported inter-class correlation coefficient (ICC) was greater than the precision results at the nurse level. This was discussed during the follow-up call and determined that reliability testing results provided did not reflect the "usual" ICC statistic but were in fact different reliability statistics that are based on the Spearman-Brown prophecy formula.

Validity

- Validity was conducted at the data element and measure score level.
- The measure was not risk adjusted
- Data element
  - The PES-NWI was developed in 2002 to measure nursing practice environments through factor analysis of 1986 survey data from staff nurses in 16 original magnet hospitals, and confirmed in 1999 data from 11,636 nurses throughout Pennsylvania (Lake, 2002). The subscales were then combined into a composite measure (Lake and Friese, 2006).
  - Results not provided but are included in the Lake 2002 paper and the Lake and Friese 2006 paper.
  - Note, NQF requires that results should be provided in the testing attachment and not only referred to in a citation.
- Measure score
  - Evaluated hypothesized relationships by computing correlation coefficients, ANOVAs, t-tests and estimating regression coefficients as summarized from two systematic reviews.
  - Developer states the results demonstrate that the measure exhibits satisfactory validity across a wide range of related constructs in many international samples across 16 years as well as in national 2015 data analyzed for measure re-endorsement.

Measures Not Discussed by the Subgroup

0729: Optimal Diabetes Care

Scientific Methods Panel Votes: Measure passes

- Reliability: H-4, M-0, L-0, I-1
- Validity: H-2, M-3, L-0, I-0
- Composite: H-1, M-4, L-0, I-0

This measure was reviewed by the Methods Panel. A summary of the measure is provided below:

Reliability

- Score-level:
  - Reliability testing done at the measure score level using the Adams' formula (signal to noise)
    - Results indicated "very good" reliability score:
    - On 2014 data, the reliability by this methodology was .908. This was repeated in 2018 for measure maintenance and the reliability was .888. The latter was performed on over 300,000 patients in 618 clinics. Reliability correlated with the number of eligible patients at the clinic with a range of .519 for those with 30 patients (the minimum per measure standard) to .994 for the largest clinics.

Validity

- Data element:
  - Data elements were validated with audit and quality checks
    - In 2018, for the diabetes measure, MNCM audited 53 medical groups; 37% of those submitting data. 89% passed the initial audit, 11% required a correction plan and all re-submitted their data and passed the audit with > 90% accuracy.
- Score-level:
  - Measure score validity was done based on clinical supposition that performance on Optimal Vascular Care measure would be similar; results of correlation testing indicated "fairly strong correlation"

3474 Hospital-level, risk-standardized payment associated with a 90-day episode of care for elective primary total hip and/or total knee arthroplasty (THA/TKA)

Scientific Methods Panel Votes: Measure passes

- Reliability: H-4, M-0, L-0, I-1
- Validity: H-2, M-2, L-0, I-0

Reliability:

- Date Element:
  - Data element reliability inferred based on standardized process for collecting claims. (This approach does not meet NQF's standard for demonstrating data element reliability).

- Score-Level:
    - Reliability testing done at the measure score level using test-retest method indicating strong reliability.
        - The agreement between the two independent assessments of each hospital was 0.931
        - The median reliability score of 0.938 calculated with 3 years of data

Validity:

- Score-Level:
    - The assessment of face validity meets NQF's requirements for demonstrating validity for new measures.
    - Data element validity was also inferred based on validity of prior (endorsed) cost measures, all which use a standard CMS approach to constructing cost and resource use measures. However, this inference does not suffice to meet NQF requirements for data element validity unless the methods and results from submission materials of these other measures are included and a summary of the analysis is provided.

Subgroup 2: Initial Evaluation Call, Thursday, October 11 from 12-2 pm ET; No Follow-Up Call Needed

During the call, Subgroup 2 discussed six measures (2539, 3366, 3470, 3443, 3445, and 3458) and accepted the preliminary analysis decisions for seven measures (3459, 3457, 3449, 3477, 3479,3480, and 3481) without further discussion. Because the subgroup was able to complete its discussion of measures during its initial call, the follow-up call scheduled for Wednesday, October 17 was canceled. The final results for the thirteen measures evaluated by Subgroup 2 are presented below.

Measures Discussed by the Subgroup

2539 Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy

Scientific Methods Panel Votes: Measure does not pass

- Reliability: H-2, M-3, L-0, I-0
- Validity: H-0, M-1, L-3, I-1

This measure was reviewed by the Methods Panel . The panel agreed that the measure, as submitted, does not meet NQF's requirements for validity because empirical validity testing was not submitted, and panel members did not consider the justification for lack of empirical validation to be sufficient.

3366 Hospital Visits after Urology Ambulatory Surgical Center Procedures

Scientific Methods Panel Votes: Measure passes

- Reliability: H-1, M-3, L-1, I-0
- Validity: H-0, M-5, L-0, I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- Reliability testing was conducted at the measure score level
- Measure score
    - Score-level reliability was demonstrated in two ways: a signal-to-noise ratio (SNR) analysis using the Adams method and a split-sample ICC (2,1).
        - NOTE that at least for the split-sample analysis, the developers did NOT limit their testing data to facilities with >=30 cases (i.e., testing aligned with specifications)
            - Split-sample (ICC (2,1) = 0.45
            - SNR (for facilities with >=30 cases) median reliability = 0.69
            - Developer notes these results indicate that there is sufficient reliability in the measure score.

Validity

- The developer provided results of an assessment of the measure's face validity.
- The developer assessed face validity in various ways; however, only the TEP assessment meets NQF's requirements for face validity.

    - Of the 14 TEP respondents, 12 (86%) indicated that they somewhat, moderately, or strongly agreed and 2 respondents moderately disagreed with the statement about whether the results from the measure can be used to differentiate poor vs good quality. They also provided reason for disagreement from one of the two who disagreed.
- Risk adjustment
    - Risk-adjustment with nine risk factors
        - The measure uses a two-level hierarchical logistic regression model to estimate ASC-level risk-standardized hospital visit rates (RSHVRs).
        - Nine risk factors:
            - 1. Age (years > 65)
            - 2. Work Relative Value Units (work RVUs)
            - 3. Benign prostatic hyperplasia with obstruction (ICD-9-CM diagnosis codes 60001, 60021, 60091; ICD-10-CM diagnosis codes N401, N403)
            - 4. Complications of specified implanted device or graft (Condition Category 176)
            - 5. Number of qualifying procedures: 1, 2, 3 or more
            - 6. Poisonings and allergic and inflammatory reactions (CC 175)
            - 7. Major symptoms, abnormalities (CC 178)
            - 8. Parkinson's and Huntington's diseases; seizure disorders and convulsions (CC 78, 79)
            - 9. Ischemic heart disease (CC 86, 87, 88, 89)
        - The Methods Panel's main concern was the decision not to include dual status in risk-adjustment approach (i.e., dual-eligible status had a statistically significant association (OR: 1.30, 95% CI: 1.13 -1.48, p = 0.0001) with the risk of a hospital visit)
        - There was also some concern with the c-statistic (0.61 in development and validation samples), which panel members found to be a bit low.
- Panel members questioned whether this measure is able to distinguish meaningful differences between providers, since an outlier analysis suggested only 19 of 1,204 ASCs were better or worse than expected.

3470 Hospital Visits after Orthopedic Ambulatory Surgical Center Procedures

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0, M-4, L-1, I-0
- Validity: H-0, M-4, L-1, I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- Reliability was assessed at the measure score level:
- Measure score:

- Score-level reliability demonstrated in two ways: Signal-to-noise ratio (SNR) using Adams method and split-sample ICC(2,1)
- NOTE that for both analyses, the developers limited their testing data to facilities with >=30 cases; this is not aligned with the measure specifications, which do not identify minimum case volume.
    - Split-sample (ICC (2,1) = 0.25 [fair agreement, according to Landis and Koch classification]
    - SNR, median reliability = 0.66 (2 years of data)
    - The developer interprets these results as indicating that there is sufficient reliability in the measure score.

Validity

- The developer assessed the face validity of the measure score.
- Developer assessed face validity in various ways; however, only the TEP assessment meets NQF's requirements for validity testing.
    - Note that developer assessed face validity by their TEP, but this does not quite meet NQF requirements for face validity.
    - NQF requires the face validity assessment to explicitly address whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
    - Of the 13 TEP respondents, 12 (92%) indicated that they somewhat, moderately, or strongly agreed with the statements above and 1 respondent moderately disagreed
        - Developers did not provide reason for disagreement
- Risk adjustment
    - Risk-adjustment with 29 risk factors
        - The measure uses a two-level hierarchical logistic regression model to estimate ASC-level risk-standardized hospital visit rates (RSHVRs).
        - 29 risk factors: age, 27 comorbidity variables, and one surgical complexity variable
        - Panel's main concern was the decision not to include dual status in risk-adjustment approach (i.e., dual-eligible status had a statistically significant association (OR: 1.35, 95% CI: 1.20 -1.52, p < 0.0001) with the risk of a hospital visit)
        - There was also some concern with the c-statistic (0.66 in development and validation samples) as panel members thought this was a bit low.
- Panel members questioned whether this measure is able to distinguish meaningful differences between providers, since outlier analysis suggests only 7 of 2,734 ASCs were better or worse than expected.

3443 All-cause emergency department utilization rate for Medicaid beneficiaries with complex care needs and high costs (BCNs)

Scientific Methods Panel Votes: Consensus Not Reached on Validity

- Reliability: H- 3, M-1, L-0, I-1

- Validity: H-0, M-3, L-1, I-1

This measure was evaluated by the Scientific Methods Panel and discussed on their call. A summary of the measure and the Panel discussion is provided below. This measure is paired with the all-cause inpatient admission rate for BCNs, referred to as "BCN-2" (measure 3445).

Reliability

- Reliability testing was conducted at the measure score level.
- Developer conducted signal-to-noise (SNR) reliability testing for BCN-1 using MAX data from 10 states.
    - Average signal-to-noise reliability estimate = 0.92 (ranging between 0.59 to 0.99 across the ten states in the sample).
        - Note: there is a small typo regarding the average - stated as 0.99 in text but shown as 0.92 in table.
    - The two states with very low sample sizes had the lowest signal-to-noise reliability estimates (i.e., 0.59 and 0.66)
- Testing was not precisely conducted for the measure as specified. Specifically, dual-eligible beneficiaries were not included in the testing due to data unavailability, but would be included in the measure if implemented.
- Panel members, in a future submission, would like to see analyses demonstrating the reliability of the data elements used in the measure. This is important because of the probable differences in the quality of Medicaid data across states.

Validity

- A face validity assessment was conducted and meets NQF requirements.
    - 10 of 17 TEP members responded to the relevant question asked as part of the face validity assessment. Seven of the 10 respondents agreed/strongly agreed that the measure can differentiate between good vs poor quality. Among those who disagreed, one misunderstood the BCN population definition and two did not give a reason.
- The risk-adjustment approach was developed using data from 10 states. The risk-adjustment model included 69 risk factors. The Panel's concerns about the risk-adjustment approach included:
    - Several risk factors are included that are neither statistically nor clinically significant. However, the risk of over-fitting was not assessed.
    - The risk-adjustment model includes a factor noted as "child'. This is confusing given the measure is limited to individuals ages 18-64.
    - Poly-pharmacy was not included as a risk-factor.
    - The developer states they did not include social risk factors due to the findings from a recent NQF report on admissions/readmissions. This is an erroneous interpretation of that report.
    - Concern with excluding prior hospital-based care utilization as a risk factor.
    - Concern around using the chronic conditions data warehouse (CCW) fields to identify comorbidities used in the risk adjustment model. However, there was no supporting literature cited to support this decision and no validation of those

variables (e.g., re-abstracting chart data) was conducted. Because this is a new measure, data element validation is not required. If endorsed, developers should consider presenting this type of analysis when the measure comes back for re-evaluation.

- Regarding other potential threats to validity:
  - Some concern regarding generalizability, given the limited dataset used to develop the measure.
  - Concern that differences in enrollment criteria between states may impact the number of beneficiaries included in the measure (i.e., particularly as the measure is specified so that Medicaid beneficiaries with <10 months of data in the 12-month lookback period are excluded from the measure). The concern is that states with differing enrollment "churn" (different populations going on and off Medicaid in any given year) may be linked to various health or social issues, which in turn could make some states artificially look better (or worse) than others.
    - Panel members recommended that the developers provide some analysis of the population that is excluded (overall and by major demographic groupings, as well as by FFS vs. managed care, for each of the 10 sates state).
  - Need for clarity regarding how missing data are handled when calculating the measure.

3445 All-cause inpatient admission rate for Medicaid beneficiaries with complex care needs and high costs (BCNs)

Scientific Methods Panel Votes: Consensus Not Reached on Validity

- Reliability: H-3 , M-1, L-1, I-0
- Validity: H-0, M-3, L-1, I-1

This measure was evaluated by the Scientific Methods Panel and discussed on their call. A summary of the measure and the Panel discussion is provided below. This measure is paired with the all-cause emergency department utilization rate for BCNs, referred to as "BCN-1" (measure 3443).

Reliability

- Reliability testing was conducted at the measure score level.
- Developer conducted signal-to-noise (SNR) reliability testing for BCN-2 using MAX data from 10 states.
  - Average signal-to-noise reliability estimate = 0.99 (ranging between 0.95 to 0.99 across the ten states in the sample).
- Testing was not precisely conducted for the measure as specified. Specifically, dual-eligible beneficiaries were not included in the testing due to data unavailability, but would be included in the measure if implemented.
- Panel members, in a future submission, would like to see analyses demonstrating the reliability of the data elements used in the measure. This is important because of the probable differences in the quality of Medicaid data across states.

Validity

- Validity was performed at the measure score level.
- Developer conducted convergent validity testing by examining the correlation between this measure and the HEDIS inpatient hospital utilization measure (IHU).
    - The IHU measure includes beneficiaries of commercial plans and Medicare rather the Medicaid population (and the specifications of the measures differed slightly).
    - There was concern about whether testing against a measure that includes different beneficiaries represents a reasonable testing approach. The question was whether the high correlation (0.82, 95%CI: (0.50, 0.94)) does help to validate the measure that assesses state Medicaid program performance or if, instead, the results merely indicate similarities in care delivery within states.
- The risk-adjustment approach was developed using data from 10 states. The risk-adjustment model included 69 risk factors. The Panel's concerns about the risk-adjustment approach included:
    - Several risk factors are included that are neither statistically nor clinically significant. However, the risk of over-fitting was not assessed.
    - The risk-adjustment model includes a factor noted as "child'. This is confusing given the measure is limited to individuals ages 18-64.
    - Poly-pharmacy was not included as a risk-factor.
    - The developer states they did not include social risk factors due to the findings from a recent NQF report on admissions/readmissions. This is an erroneous interpretation of that report.
    - Concern with excluding prior hospital-based care utilization as a risk factor.
    - Concern around using the chronic conditions data warehouse (CCW) fields to identify comorbidities used in the risk adjustment model. However, there was no supporting literature cited to support this decision and no validation of those variables (e.g., re-abstracting chart data) was conducted. Because this is a new measure, data element validation is not required. If endorsed, developers should consider presenting this type of analysis when the measure comes back for re-evaluation.
- Regarding other potential threats to validity:
    - Some concern regarding generalizability, given the limited dataset used to develop the measure.
    - Concern that differences in enrollment criteria between states may impact the number of beneficiaries included in the measure (i.e., particularly as the measure is specified so that Medicaid beneficiaries with <10 months of data in the 12-month lookback period are excluded from the measure). The concern is that states with differing enrollment "churn" (different populations going on and off Medicaid in any given year) may be linked to various health or social issues, which in turn could make some states artificially look better (or worse) than others.

- Panel members recommended that the developers provide some analysis of the population that is excluded (overall and by major demographic groupings, as well as by FFS vs. managed care, for each of the 10 sates state).

3458 Successful Transition after Long-Term Institutional Stay

Scientific Methods Panel Votes: Measure does not pass on Validity

- Reliability: H-0, M-2, L-3, I-0
- Validity: H-0, M-0, L-5, I-0

This measure was reviewed by NQF's Scientific Methods Panel. The panel agreed that the measure, as submitted, does not meet NQF's requirements for validity due to concerns with the risk-adjustment approach that was developed and validated using a very small sample.

Measures Not Discussed by the Subgroup

3456 Admission to an Institution from the Community

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0, M-3, L-1, I-0
- Validity: H-1, M-2, L-1, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below:

Reliability

- Reliability was assessed at the measure score level
- Some Panel members were concerned about the measure's exclusion of the month that an enrollee dies, and any subsequent months of enrollment, from the denominator. Panel members suggested greater clarity in the specification of this exclusion is needed.
- Measure score
  - Signal-to-noise analysis using the Morris method conducted
  - HPPLs with 10 or fewer outcome events (i.e. admissions) were excluded based on CMS standards; note that this testing is not aligned with the measure specifications, which do not identify a minimum case volume.
  - Results:

Signal-to-noise reliability of age-stratified rates

Measure	Level of aggregation	Average reliability score	Interquartile range of reliability scores	Median reliability score	% Plans exceeding 0.4 SNR
Plans with 11 or more numerator events					
Short-stay, 18-64	12 HPPLs	0.5322	0.3082 – 0.8156	0.4961	50%
Short-stay, 65-74	11 HPPLs	0.5491	0.4288 – 0.9219	0.5707	55%
Short-stay, 75-84	10 HPPLs	0.7238	0.7179 – 0.7777	0.7448	80%
Short-stay, 85+	9 HPPLs	0.7353	0.7011 – 0.7832	0.7625	78%
Medium-stay: 18-64	13 HPPLs	0.6060	0.3535 – 0.8764	0.6293	62%
Medium-stay, 65-74	12 HPPLs	0.8279	0.7605 – 0.9552	0.8705	92%
Medium-stay, 75-84	14 HPPLs	0.8975	0.8506 – 0.9828	0.9249	100%
Medium-stay, 85+	11 HPPLs	0.8846	0.9166 – 0.9880	0.9593	88%
Long-stay: 18-64	12 HPPLs	0.9417	0.9195 – 0.9918	0.9668	100%
Long-stay, 65-74	12 HPPLs	0.9328	0.9314 – 0.9928	0.9565	100%
Long-stay, 75-84	12 HPPLs	0.9399	0.9604 – 0.9984	0.9827	100%
Long-stay, 85+	12 HPPLs	0.9093	0.9568 – 0.9988	0.9941	100%

Source: Mathematica analysis of data from four MLTSS plans and fourteen health plan product lines

- Average and median SNRs reported for the 12 stay/age groupings
  - Averages range between 0.53 to 0.94
  - Medians range between 0.50 to 0.99
  - Highest estimates for long stays, regardless of age group
  - Could be problematic for short stays for those under age 75

Validity

- Validity was assessed at the measure score level.
- Measure score
  - The developer conducted score-level validity testing by comparing measure results with results from two other measures (3457: Minimizing institutional length of stay and 3458: Successful transition after long-term institutional stay). They assessed convergent validity based on the Spearman correlation.
  - Results generally were as hypothesized:
    - Their analyses found that the twelve rates that comprise the *Admission to an institution from the community* measure generally had positive, significant associations with one another and relationships tended to be stronger within age strata and more similar rates (i.e., rates for a particular age category were aligned across stay types, rates for a particular stay type were aligned across ages within the stay type, and

short-stay results were more aligned with medium-stay results than long-stay results).
        - They saw a moderate, negative correlation between the long-stay rates on this measure and performance on a measure of minimizing institutional length of stay.
        - They observed mostly negative relationships between the twelve *Admission to an institution from the community* measure rates and the *Successful transition after long-term institutional stay* measure rates among the 14 HPPLs; this finding is expected because while the *Admission to an institution from the community* measure captures a less desirable outcome (institutional admission), the *Successful transition after long-term institutional stay* rates reflect positive outcomes after institutional discharge.
- Risk adjustment
    - Risk adjustment via stratification of four age groups
    - No social risk factors included
        - Note: Developer cites NQF's SES trial results as rationale for not including SES. NQF does not agree with this interpretation.
    - The panel noted concern about a lack of compelling analysis to support the exclusion of dual status in the risk adjustment model, and were also concerned that there may not be enough events to allow for adequate risk-adjustment.
- Panel members also noted concerns about:
    - whether the measure as specified truly reflects quality (particularly given that different plans may score differently based on the stay/age groupings).
    - low event rates and potential difficulty demonstrating meaningful differences between plans

3457 Minimizing Institutional Length of Stay

Scientific Methods Panel Votes: Measure passes

- Reliability: H-2, M-1, L-1, I-0
- Validity: H-2, M-1, L-1, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below:

Reliability

- Reliability was assessed at the measure score level.
- Measure score
    - Signal-to-noise analysis using the Morris method
    - Results:

Signal-To-Noise Reliability [a]

HPPL	Mean SNR
HPPL-01	0.9511
HPPL-02	0.8454

HPPL	Mean SNR
HPPL-03	0.6326
HPPL-04	0.9907
HPPL-05	0.9799
HPPL-06	0.9237
HPPL-07	0.9861
HPPL-08	0.8592
HPPL-09	0.9309
HPPL-11	0.9809
HPPL-12	N/A
HPPL-13	0.9862
HPPL-14	0.8736
HPPL-17	0.9870
Average	**0.9175**

Source: Mathematica analysis of data from four MLTSS plans and fourteen health plan product lines.
[a] SNR values excluding plans with fewer than 11 enrollees in the numerator.

- Average and plan-level SNRs reported: Plan SNRs ranged 0.63 to 0.99

Validity

- Validity was assessed at the measure score level.
- Measure score
  - The developer conducted score-level validity testing (construct validity) by comparing measure results with results from two other measures (3456: Admission to an institution from the community and 3458: Successful transition after long-term institutional stay).
  - Results
    - Results generally were as hypothesized.

Results of Correlation Analyses

Measure	Strata	Correlation Coefficient	p-value
Admission to an institution from the community: Long Stay (101+ days)**	Age 18-64	-0.12321	0.6747
	Age 65-74	-0.54125*	0.0456
	Age 75-84	-0.47965	0.0826
	Age 85+	-0.56292*	0.0361
Successful transition after long-term (101+ days) institutional stay	Risk-adjusted for age, gender and co-morbid conditions	0.89091**	0.0005

Source: Mathematica analysis of data from four MLTSS plans and fourteen health plan product lines.
*Correlation was significant at p<.05

**Correlations between the proposed measure and Long-stay admissions was hypothesized to be negative because the measure of long-stay admissions is a "lower is better" measure.

        - They saw a moderate, negative correlation between a measure of utilization of long-stay institutional care and performance on this measure of minimizing institutional length of stay.
        - They saw an even stronger positive correlation between a measure of successful transition to the community after long-stay institutional stay and this measure of minimizing institutional length of stay (correlation=0.89, p = 0.0005).
- Risk adjustment
    - Risk adjustment with 18 factors
        - Risk Model: A logistic regression was used to model the log-odds of an admission being successfully discharged within 100 days of admission.
            - Note: The panel members suggest that Standing Committee weigh in on the 100-day threshold; they were unsure from a content perspective if this threshold was appropriate
        - Risk Factors: Age, gender, dual-eligibility for Medicaid and Medicare, hospital utilization in the period prior to admission to the institutional facility, enrollment in the MLTSS plan for 6 months, and comorbid conditions (Alzheimer's disease and related disorders, asthma, intellectual disabilities, mental health conditions, stroke).
    - Developers noted a conceptual rationale for marital status and functional status, but cited lack of data as reason for not analyzing these variables.
    - No social risk factors included
        - Note: Developer cites NQF's SES trial results as rationale for not including SES. NQF does not agree with this interpretation.
- Panel members also noted concern of a lack of compelling analysis to support the exclusion of dual status in the risk adjustment model

3449 Hospitalization for Ambulatory Care Sensitive Conditions for Dual Eligible Beneficiaries

Scientific Methods Panel Votes: Measure passes

- Reliability: H-4 , M-0, L-0, I-0
- Validity: H-3 , M-1, L-0, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure and the Panel discussion is provided below.

Reliability

- Reliability testing was conducted at the measure score level (required, as this is a composite measure).
- Reliability testing was conducted using a signal to noise analysis (using a nonparametric method developed by Morris) to evaluate reliability for each composite rate for each stratum: (1) community-dwelling home and community-based services (HCBS) users, (2)

community-dwelling non-HCBS users, or (3) non-community-dwelling (institutionalized) population.

- Data used for testing obtained from all 50 states + DC (October 2014 – September 2015)
    - *Community-dwelling HCBS stratum*: Mean reliability >0.89 for the acute, chronic, and total groups (ranging between 0.48-0.99)
    - *Community-dwelling non-HCBS stratum*: Mean reliability >0.94 for the acute, chronic, and total groups (ranging between 0.71-0.99)
    - *Institutionalized stratum:* Mean reliability >0.86 for the acute, chronic, and total groups (ranging between 0.34-0.99)

Validity

- Empirical validity testing of both the overall composite measure score and the component measure scores was conducted.
- Calculated the Spearman rank correlation between each rate (acute, chronic, total) for each stratification (HCBS, non-HCBS, institutionalized—a "within measure" analysis), and for selected rates/strata with four other measures (a similar dual-eligible FFS HCBS measure of hospitalization for ambulatory sensitive conditions and Medicare FFS readmission measures for AMI, heart failure, and COPD). Developers hypothesized that states that perform well on one rate (acute, chronic, and composite) are likely to perform well on the other rates, particularly for similar rates across each strata of beneficiaries (HCBS, non-HCBS, institutionalized).
    - *Within measure rate correlations:* Correlations ranged from 0.19 to 0.93, although most can be classified as moderate (i.e., between 0.25 and .075). These results for the most part supported the developers' hypotheses.
    - *FFS dual eligible HCBS Ambulatory Care Sensitive Condition (ACSC) measure*:
        - Correlations ranged from 0.15 to 0.69, although most can be classified as moderate (i.e., between 0.25 and .075). These results for the most part supported the developers' hypotheses.
    - *Medicare FFS readmission measures*:
        - AMI: Correlations ranged from 0.17 to 0.71, with the weakest between the overall composite score and the readmission score in the institutionalized stratum.
        - Heart failure: Correlations ranged from 0.25 to 0.67, with the weakest between the overall composite score and the readmission score in the institutionalized stratum.
        - COPD: Correlations ranged from 0.25 to 0.71, with the weakest between the overall composite score and the readmission score in the institutionalized stratum.
        - These results for the most part supported the developers' hypotheses.
    - Calculated the Spearman rank correlation between each component rate (acute and chronic) with the 10 components of a similar dual-eligible FFS HCBS measure of hospitalization for ambulatory sensitive conditions and with two other measures of hospitalization (for cellulitis and pressure ulcer). This analysis was NOT conducted for each stratum separately, and therefore does not represent testing for the measure as specified.

- The risk-adjustment models included 95 risk factors for the acute component measure, 83 risk factors for the chronic component measure, and 106 risk factors for the overall composite measure**.**
    - The modeling methodology employed a two-step design, first using logistic regression to model the log-odds of having any qualifying ACSC admission during the measurement period, and the second using Poisson regression to model the total count of qualifying ACSC admissions experienced over the measurement period.
    - The developers did provide a conceptual rationale regarding the linkage between social risk factors and the measured outcome.
    - The developer states they did not include social risk factors due to the findings from a recent NQF report on admissions/readmissions. This is an erroneous interpretation of that report.
    - Model discrimination for stage one of the model was analyzed via the c-statistic. Values ranged from 0.661 to 0.851 in the development sample and from 0.661 to 0.854 in the validation sample.
    - To examine calibration of the modeling approach, developers developed risk-decile plots to compare observed vs. predicted values across rates/strata and also calculated observed-to-predicted ratios for various subgroup populations across rates/strata. Developers interpreted the results as demonstrating that the risk models are well-calibrated.
    - There was some concern by one Methods Panel member regarding potential overfitting of the model, as many of the clinical factors have odds-ratios incidence rate ratios with 95% confidence intervals that include 1.0.
- Concerns regarding potential threats to validity:
    - There was some concern that the developers did not report an analysis of missing data (although developers did that records were excluded from the measure if data elements are missing and stated that <100 records had missing values for state).

Composite

- The developer used the Cronbach's alpha statistic to assess internal consistency of the measure components. However, these were not calculated separately by rate/strata. Values ranged from 0.69 to 0.82. The developer also presented observed rates and overall percentages for each of the individual components that formed the acute and chronic components of the measure, although this was done at the state level rather than by strata.

3477 Discharge to Community-Post Acute Care Measure for Home Health Agencies (HHA)

Scientific Methods Panel Votes: Measure passes

- Reliability: H-4, M-0, L-0, I-0
- Validity: H-2, M-2, L-0, I-0

This measure was reviewed by the Methods Panel. A summary of the measure is provided below:

Reliability

- Data element:
  - The materials submitted did not meet NQF's requirements for data element testing. However, measure score testing was provided so this measure was still determined to be reliable.
    - The developer states that data element reliability is inferred based on the assumption of inherent accuracy of Medicare claims which are used for reimbursement and the use of claims in other NQF-endorsed measures. However, this inference does not suffice to meet NQF requirements for data element validity unless the methods and results from submission materials of these other measures are included and a summary of the analysis is provided.
    - The developer also stated auditing programs are used to assess accuracy of claims data, however, did not provide the results. While auditing is an appropriate technique to assess data element accuracy, this does not meet NQF requirements for data element testing if results of the audit are not provided.
- Score-level:
  - Score-level reliability was conducted using split sample ICC and signal to noise approach; results indicated *"excellent"* ICC's between samples and "*good to excellent*" mean signal to noise ratios.
  - Method of signal-to-noise analysis is different from what NQF usually receives (i.e., rate divided by width of 95% CI of rate) but it is acceptable.

Validity

- Data element:
  - Data element testing was done comparing two-gold standard data element authoritative sources for the discharge to community setting variable using home health claims and OASIS assessment which showed a high rate of agreement.
  - The developers also performed known group testing which showed agreement in the direction of the relationship between specific patient characteristics and the performance rates of the measure.
- Score-level:
  - Measure score validity was demonstrated by testing whether a facility's performance and percentile rank on the successful discharge to community measure was correlated with its performance and percentile rank on five claims-based measures.
  - Face validity was only conducted on the measure concept and results of systematic collection of input was not provided.
  - Concern from many that the risk adjustment model did not include dual status although it was statistically significant in the model

3479 Discharge to Community-Post Acute Care Measure for Inpatient Rehabilitation Facilities (IRF)

Scientific Methods Panel Votes: Measure passes

- Reliability: H-3, M-1, L-0, I-1
- Validity: H-2, M-2, L-0, I-1

This measure was reviewed by the Methods Panel. A summary of the measure is provided below:

Reliability

- Data element:
  - Did not meet NQF's requirements for data element testing. However, measure score testing was provided so this measure was still determined to be reliable.
    - The developer states that data element reliability is inferred based on the assumption of inherent accuracy of Medicare claims which are used for reimbursement and the use of claims in other NQF-endorsed measures. However, this inference does not suffice to meet NQF requirements for data element validity unless the methods and results from submission materials of these other measures are included and a summary of the analysis is provided.
    - The developer also stated auditing programs are used to assess accuracy of claims data, however, did not provide the results. While auditing is an appropriate technique to assess data element accuracy, this does not meet NQF requirements for data element testing if results of the audit are not provided.
- Score-level:
  - Score-level reliability was conducted using split sample ICC and signal to noise approach; results indicated *"excellent"* ICC's between samples and "*good to excellent*" mean signal to noise ratios.
  - Method of signal-to-noise analysis is different from what NQF usually receives (i.e., rate divided by width of 95% CI of rate) but still acceptable.

Validity

- Data element:
  - Data element testing was done comparing two gold standard authoritative data sources for discharge to community setting using IRF-PAI (Patient Assessment Instrument) and IRF claims data which showed a high rate of agreement.
- Score-level:
  - Measure score validity was demonstrated by testing whether a facility's performance and percentile rank on the successful discharge to community measure was correlated with its performance and percentile rank on three claims-based measures.
  - Face validity was only conducted on the measure concept and results of systematic collection of input was not provided.
  - Concern from many that the risk adjustment model did not include dual status although it was statistically significant in the model

3480 Discharge to Community-Post Acute Care Measure for Long-Term Care Hospitals (LTCH)

Scientific Methods Panel Votes: Measure passes

- Reliability: H-1, M-3, L-0, I-1
- Validity: H-2, M-2, L-0, I-1

This measure was reviewed by the Methods Panel. A summary of the measure is provided below:

Reliability

- Data element:
  - Did not meet NQF's requirements for data element testing. However, measure score testing was provided so this measure was still determined to be reliable.
    - The developer states that data element reliability is inferred based on the assumption of inherent accuracy of Medicare claims which are used for reimbursement and the use of claims in other NQF-endorsed measures. However, this inference does not suffice to meet NQF requirements for data element validity unless the methods and results from submission materials of these other measures are included and a summary of the analysis is provided.
    - The developer also stated auditing programs are used to assess accuracy of claims data, however, did not provide the results. While auditing is an appropriate technique to assess data element accuracy, this does not meet NQF requirements for data element testing if results of the audit are not provided.
- Score-level:
  - Score-level reliability was conducted using split sample ICC and signal to noise approach; results indicated *"excellent"* ICC's between samples and "*good to excellent*" mean signal to noise ratios.
  - Method of signal-to-noise analysis is different from what NQF usually receives (i.e., rate divided by width of 95% CI of rate) but is acceptable.

Validity

- Data element:
  - Data element testing was done comparing to a gold standard authoritative data source for discharge to community setting using LTCH-CARE data and IRF claims data which showed a high rate of agreement.
  - The developers also performed known group testing which showed agreement in the direction of the relationship between specific patient characteristics and the performance rates of the measure.
- Score-level:
  - Measure score validity was demonstrated by testing whether a facility's performance and percentile rank on the successful discharge to community measure was correlated with its performance and percentile rank on 3 claims-based measures. The testing approach and results were determined to be adequate by the Panel.

- Face validity was only conducted on the measure concept and results of systematic collection of input was not provided.
- Concern from many that the risk adjustment model did not include dual status although it was statistically significant in the model.

3481 Discharge to Community-Post Acute Care Measure for Skilled Nursing Facilities (SNF)

Scientific Methods Panel Votes: Measure passes

- Reliability: H-1, M-3, L-0, I-1
- Validity: H-2, M-2, L-0, I-1

This measure was reviewed by the Methods Panel. A summary of the measure is provided below:

Reliability

- Data element:
  - Did not meet NQF's requirements for data element testing. However, measure score testing was provided so this measure was still determined to be reliable.
    - The developer states that data element reliability is inferred based on the assumption of inherent accuracy of Medicare claims which are used for reimbursement and the use of claims in other NQF-endorsed measures. However, this inference does not suffice to meet NQF requirements for data element validity unless the methods and results from submission materials of these other measures are included and a summary of the analysis is provided.
    - The developer also stated auditing programs are used to assess accuracy of claims data, however, did not provide the results. While auditing is an appropriate technique to assess data element accuracy, this does not meet NQF requirements for data element testing if results of the audit are not provided.
- Score-level:
  - Score-level reliability was conducted using split sample ICC and signal to noise approach; results indicated *"excellent"* ICC's between samples and "*good to excellent*" mean signal to noise ratios.
  - Method of signal-to-noise analysis is different from what NQF usually receives (i.e., rate divided by width of 95% CI of rate) but is acceptable.

Validity

- Data element:
  - Data element testing was done comparing two-gold standard authoritative data sources for discharge to community setting using variables/data elements within MDS discharge assessments and SNF data which showed high rates of agreement.
  - The developers also performed known group testing which showed agreement in the direction of the relationship between specific patient characteristics and the performance rates of the measure.
- Score-level:

- Measure score validity was demonstrated the correlation between performance scores of this measure to two claims-based measures and to six short-stay measures which showed correlation.
- Face validity was only conducted on the measure concept and results of systematic collection of input was not provided.
- Concern from many that the risk adjustment model did not include dual status although it was statistically significant in the model, particularly for SNFs with higher proportions of dual eligible patients.
- The large number of exclusions identified for the measure may require two years of data for each institution in order to accrue sufficient numbers to perform the computations as specified. There was concern that a two-year lag does not provide timely information to the health care provider regarding any changes in practice that were implemented to improve patient discharge rates.

Subgroup 3: Initial Evaluation Call, Friday, October 12 from 2-4 pm ET; Follow-Up Evaluation Call, Thursday, October 18 from 2-4 pm ET

During its initial call, Subgroup 3 discussed six measures (3309, 0964, 2936, 3478, 2561 and 2563) and accepted the preliminary analysis decisions for two measures (2377 and 2459) without further discussion. The subgroup discussed two additional measures (3483 and 3484) during the follow-up call. The final results for the ten measures evaluated by Subgroup 3 are presented below.

Measures Discussed by the Subgroup

3309 Risk-Standardized Survival Rate (RSSR) for In-Hospital Cardiac Arrest

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0, M-5, L-0, I-0
- Validity: H-0, M-5, L-0, I-0

This measure was reviewed by the Scientific Methods Panel and discussed on the call. A summary of the measure and the Panel discussion is provided below.

Reliability

- Reliability testing was conducted at the measure score level using a signal-to-noise (SNR) analysis (specifically, Adams' beta-binomial method)
    - Testing data
        - 2011-2015 analysis:
            - 326 hospitals included; total of 61,934 cardiac arrest events and 14,782 (23.9%) patients survived to hospital discharge
            - Average number of quality reporting events per hospital =190
            - Range of quality reporting events was 1 to 1222
            - Range for number of patients surviving to hospital discharge was 0 to 344.
        - 2013 analysis
            - 273 hospitals included; total of 17,992 cardiac arrest events and 4417 (24.5%) patients survived to hospital discharge
            - Average number of quality reporting events per hospital =66
            - Range of quality reporting events was 1 to 360
            - Range for number of patients surviving to hospital discharge was 0 to 121
            - Analysis restricted the analyses to the 206 hospitals that had a minimum number of 20 quality reporting events
        - 2014 analysis
            - 259 hospitals included; total of 17,244 cardiac arrest events and 4163 (24.1%) patients survived to hospital discharge
            - Average number of quality reporting events per hospital =67
            - Range of quality reporting events was 1 to 409

- Range for number of patients surviving to hospital discharge was 0 to 124
- Analysis restricted the analyses to the 200 hospitals that had a minimum number of 20 quality reporting events

- Results:
  - Using entire prospective validation period (2011-2015): SNR mean= 0.76; median= 0.78
  - 2013: SNR mean = 0.70; median = 0.72
  - 2014: SNR mean = 0.67; median = 0.68
- Panelists expressed desire for more than just means and medians

Validity

- The developer assessed the face validity of the measure score.
  - The face validity assessment adheres to NQF requirements
  - 34 of 35 TEP members responded
  - Mean rating = 3.76 (out of 5)
  - 71% of respondents (n=24) either agreed or strongly agreed with the following statement "*The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*".
  - 18% of respondents (n=6) neither agreed or disagreed
    - Some of these respondents had no expertise in risk-adjustment/clinical matters; one was concerned that measure doesn't account for DNR rates across hospitals
  - 12% of respondents (n=4) disagreed or strongly disagreed
    - Concerns with lack of adjustment for DNR or other risk factors (e.g., race)
- Risk adjustment
  - Risk adjusted with 9 risk factors
    - Developers provided no conceptual rationale regarding potential relationships between social risk factors and the outcome of interest (survival after heart attack), other than noting that clinicians responding to in-hospital cardiac arrest would not be aware of a patient's social or economic risk.
    - Developers did not include social risk factors in the adjustment approach.
  - Discrimination statistics (c-statistics)
    - Initial sample: 0.704
    - 2012: 0.694
    - 2013: 0.709
    - 2014: 0.703
    - 2011-2015: 0.706
  - Model calibration assessed via examination of plots of observed versus predicted values (graphs presented, along with R^2 statistics)

    - 2012: 0.992
    - 2013: 0.992
    - 2014: 0.990
    - 2011-2015: 0.997
- Some panel members expressed concern that DNR (Do Not Resuscitate) status is not accounted for in the measure

0964 Therapy with aspirin, P2Y12 inhibitor, and statin at discharge following PCI in eligible patients

Scientific Methods Panel Votes: Consensus not reached on Validity

- Reliability: H-3 , M-1, L-1, I-0
- Validity: H-0, M-3, L-2, I-0
- Composite: H-2 , M-2 , L-1 , I-0

This measure was reviewed by the Scientific Methods Panel and discussed on the call. A summary of the measure and the Panel discussion is provided below.

Reliability

- Reliability testing was conducted at the measure score level using a split-sample methodology
    - Results: Pearson correlation: r=0.90

Validity

- Empirical validity testing was conducted at the measure score level. Developers also described the conduct of a face validity assessment; however, that assessment does not conform to NQF's requirements.
- Developers conducted a construct validation analysis by correlating the results of this measure with results from two measures of 30-day all-cause mortality following PCI (NQF 0536, which includes patients with STEMI/shock, and NQF 0535, which includes patients without STEMI/shock) using data from Q4 2013 to Q3 2014.
    - Developers hypothesized that providing discharge medications for PCI patients leads to better short-term outcomes.
    - Results
        - Pearson correlation coefficient between this measure and STEMI/Shock mortality measure (NQF: 0536): -0.07465 (n=1,273)
        - Pearson correlation coefficient between this measure and NSTEMI/No Shock mortality measure (NQF: 0535): -0.16380 (n=1,283)
        - These results supported the developers' hypothesis (i.e., better provision of discharge medications was associated with lower mortality), although the magnitude of the correlations was low.
    - Panel members were concerned about the low, albeit statistically significant correlations results. They applauded the effort to assess the association with a relevant outcome, but questioned whether mortality was the best outcome to

investigate. They suggested that a more proximal outcome measure may have been more suitable.

- Additional concerns regarding validity
  - The Panel noted the overall high performance rates across facilities and questioned whether meaningful differences exist (Mean=93.58; median=95.83; 25th percentile=91.87)
  - The denominator of the P2Y12 inhibitor component is quite a bit narrower than that of the other two components (i.e., restricted to patients undergoing PCI with stenting but no contraindication to the P2Y12 inhibitor). The concern is that facility performance may be impacted not only by the performance on the components in the measure, but also on the relative frequency of PCI with and without stenting.
  - Panel members noted that hospitals that do not pass the data quality review for the NCDR registry are not included in the measure.

Composite

- Developers computed hospital-level results for the three components and correlated them with the composite results (via the Pearson correlation statistic)
  - Aspirin: r=0.7774
  - P2Y12: r=0.5910
  - Statin: r=0.9508
- Panel members would have liked to have seen more analysis to support the equal-weighting decision
- Panel members expressed concern about the utility of including the aspirin and P2Y12 components in the composite

2936 Admissions and Emergency Department (ED) Visits for Patients Receiving Outpatient Chemotherapy

Scientific Methods Panel Votes: Measure passes

- Reliability: H-1 , M-4, L-0, I-0
- Validity: H-1, M-4, L-0, I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- Reliability was conducted at the measure score-level.
  - Score-level reliability was demonstrated in two ways: Signal-to-noise (SNR) ratio using Adams method and via a split-sample ICC (2,1)
  - NOTE: For both signal-to-noise and split-sample, testing was limited to hospitals with at least 25 and 50 patients, respectively. As such, testing was not consistent with the measure's specifications.
  - **Signal-to-noise results:**
    - Cancer hospitals (n=11): Admissions measure median reliability=0.7848; ED measure median reliability=0.9808

        - Non-cancer hospitals (n=1,524): Admissions measure median reliability=0. 6027; ED measure median reliability=0. 7326
    - **Split-sample results:**
        - Cancer hospitals (n=11): Admissions measure ICC=0.6704; ED measure ICC=0.8904
        - Non-cancer hospitals (n=1,099): Admissions measure ICC=0. 4314; ED measure ICC=0. 3585

Validity

- Validity was conducted via face validity.
- The developer conducted various assessments to demonstrate face validity. Only the assessment by the 2018 Expert Workgroup (EWG) meets NQF's requirements for face validity.
- The Methods Panel had questions about the risk-adjustment approach, which NQF staff clarified based on the testing results (e.g. the definition of concurrent radiology). Staff highlighted the developer's extensive analysis and discussion about consideration of social risk factors and reiterated that the inclusion, or lack of, specific risk-factors should not be a reason to reject a measure.
- **Meaningful differences results:**
    - Cancer hospitals (n=11), admission measure: 1 identified as performing significantly better than the national rate
    - Cancer hospitals (n=11), ED measure: 3 identified as performing significantly better than the national rate; 3 identified as performing significantly worse than the national rate
    - Non-cancer hospitals (n=3,562), admission measure: 13 identified as performing significantly better than the national rate; 65 identified as performing significantly worse than the national rate
    - Non-cancer hospitals (n=3,562), ED measure: 26 identified as performing significantly better than the national rate; 33 identified as performing significantly worse than the national rate

3478 Surgical Treatment Complications for Localized Prostate Cancer

Scientific Methods Panel Votes: Measure does not pass

- Reliability: H-0, M-2, L-4, I-0
- Validity: H-0, M-3, L-2, I-1

This measure was reviewed by the Methods Panel and did not pass. Methods Panel members expressed a desire to see a more detailed explanation of the methodology by which the measure results are calculated; Panel members also felt that, given the developer's recommendation that measure results be stratified by procedure type for reporting purposes, separate reliability testing should be provided for the stratified results, particularly because there were concerns about lower reliability in hospitals with small case volumes. In addition, Panel members wanted to see a more detailed justification for the lack of risk adjustment A summary is provided below:

2561 STS Aortic Valve Replacement (AVR) Composite Score

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0, M-5, L-0, I-0
- Validity: H-0, M-4, L-0, I-2
- Composite: H-3, M-2, L-0, I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- Score-level:
    - Reliability testing was completed at the score level using signal to noise ratio
        - Posterior mean of reliability= 0.49
        - The posterior median, lower and upper boundaries of 95% credible intervals: 0.49 (0.44, 0.54)
        - The posterior mean, lower and upper boundaries of 95% credible intervals for participants with 50+ operations (n = 534): 0.59 (0.54, 0.64)
        - The posterior mean, lower and upper boundaries of 95% credible intervals for participants with 100+ operations (n = 264): 0.69 (0.63, 0.75)

Validity

- Score-level:
    - Face validity assessment of the composite
        - The developers reported completion of a face validity assessment, however, as submitted, it did not meet NQF requirements and therefore could not be considered when rating for validity.
            - NQF requirements for face validity: 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.
    - Predictive validity (stability over time) of the composite
        - STS participants were labeled as “better than average outliers” (3 Stars) if it was at least 95% certain that the participant’s true composite score was better than the overall STS average composite score. Participants were labeled as “worse than average outliers” (1 Star) if it was at least 95% certain the participant’s true composite score was worse than the overall STS average composite score.
        - The developers’ analysis, described as predictive validity, examined stability of star ratings over a 3 year period. The greatest stability was found among those with 2-star ratings.
    - Content validity of the components

        - The developers' content validity testing sought to demonstrate the components of the composite represent quality aortic valve replacement. Their approach assessed morbidity and mortality results for those providers classified as 1-star, 2-star, and 3-star based on the composite measure results
    - During the call, subgroup members expressed concern about the score-level testing methodology, noting that the star-rating consistency over time is expected and is not an appropriate approach to demonstrating validity. They also questioned the utility of the content validation approach for this measure.
    - Subgroup members also questioned the approach used to determine the inclusion of SDS factors in risk adjustment model, although NQF staff reminded the subgroup members that inclusion, or lack of, specific risk-factors should not be a reason to reject a measure.
        - The developers describe that race was included as a "genetic factor" as it relates to effects of medication efficacy and prevalence of certain diseases like diabetes and hypertension, rather than being considered social factor.

Composite Construction

- The correlation of each domain-specific estimate with the overall composite score was calculated:
    - Pearson correlation for mortality component=0.31
    - Pearson correlation for mortality component=0.77
- The correlation of the morbidity components with the morbidity estimate was calculated"
    - Pearson correlation values ranged from 0.10 to 0.87
    - Spearman correlation values ranged from 0.05 to 0.83

2563 STS Aortic Valve Replacement (AVR) + Coronary Artery Bypass Graft (CABG) Composite Score

Scientific Methods Panel Votes: Measure passes

- Reliability: H-1, M-4, L-0, I-0
- Validity: H-0, M-4, L-0, I-2
- Composite: H-3, M-2, L-0, I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- Score-level:
    - Reliability testing was completed at the score level using signal to noise ratio
        - Posterior mean of reliability= 0.50
        - The posterior median, lower and upper boundaries of 95% credible intervals: 0.50 (0.45, 0.55)
        - The posterior mean, lower and upper boundaries of 95% credible intervals for participants with 50+ operations (n = 372): 0.62 (0.56, 0.67)

        - The posterior mean, lower and upper boundaries of 95% credible intervals for participants with 100+ operations (n = 138): 0.67 (0.59, 0.74)

Validity

- Score-level:
    - Face validity assessment of the composite
        - The developers reported completion of a face validity assessment, however, as submitted, it did not meet NQF requirements and therefore could not be considered when rating for validity.
            - NQF requirements for face validity: 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.
    - Predictive validity (stability over time) of the composite
        - STS participants were labeled as "better than average outliers" (3 Stars) if it was at least 95% certain that the participant's true composite score was better than the overall STS average composite score. Participants were labeled as "worse than average outliers" (1 Star) if it was at least 95% certain the participant's true composite score was worse than the overall STS average composite score.
        - Developers' analysis described as predictive validity examined stability of star ratings over a 3 year period. The greatest stability was found among those with 2-star ratings.
    - Content validity of the components
        - The developers' content validity testing sought to demonstrate the components of the composite represent quality aortic valve replacement. Their approach assessed morbidity and mortality results for those providers classified as 1-star, 2-star, and 3-star based on the composite measure results
    - During the call, subgroup members expressed concern about the score-level testing methodology, noting that the star-rating consistency over time is expected and is not an appropriate approach to demonstrating validity. They also questioned the utility of the content validation approach for this measure.
    - Subgroup members also questioned the approach used to determine the inclusion of SDS factors in risk adjustment model, although NQF staff reminded the subgroup members that inclusion, or lack of, specific risk-factors should not be a reason to reject a measure.
        - The developers describe that race was included as a "genetic factor" as it relates to effects of medication efficacy and prevalence of certain

diseases like diabetes and hypertension, rather than being considered social factor.

Composite Construction

- The correlation of each domain-specific estimate with the overall composite score was calculated:
  - Pearson correlation for mortality component=0.31
  - Pearson correlation for morbidity domain score=0.66
- The correlation of the morbidity components with the morbidity estimate was calculated:
  - Pearson correlation values ranged from 0.16 to 0.90
  - Spearman correlation values ranged from 0.11 to 0.88

3483 Adult Immunization Status

Scientific Methods Panel Votes: Measure did not pass on reliability and validity

- Reliability: H-0, M-0, L-4, I-1
- Validity: H-0, M-1, L-3, I-1
- Composite: H-3, M-2, L-0, I-0

This measure was reviewed by the Methods Panel and did not pass. A summary is provided below:

Reliability

- There were concerns regarding clarity of the specifications including, the weighting of components, description of units of measurement in numerator and denominator, and continuity of enrollment
- There were concerns with the "perfect" range (1.0) of the reliability score and whether calculations were correctly applied.

Validity

- Face validity was assessed but did not conform to NQF requirements
- Validity testing was performed using data from only 3 health plans. There were concerns that this is not a sufficient sample size.
- There were concerns whether the results of the correlation analysis performed to demonstrate score-level validity were correctly calculated. These correlations were used to demonstrate construct validity between the composite measure and its components and the composite and other similar immunization measures.

3484 Prenatal Immunization Status

Scientific Methods Panel Votes: Measure did not pass on reliability and validity

- Reliability: H-0, M-1, L-3, I-1
- Validity: H-0, M-1, L-2, I-2
- Composite: H-4, M-0, L-0, I-1

This measure was reviewed by the Methods Panel and did not pass. A summary is provided below:

Reliability

- There were concerns regarding clarity of the specifications including, the weighting of components, description of units of measurement in numerator and denominator, and continuity of enrollment
- There were concerns with the "perfect" range (1.0) of the reliability score and whether calculations were correctly applied.

Validity

- Face validity was assessed but did not conform to NQF requirements
- Validity testing was performed using data from only 3 health plans. There were concerns that this is not a sufficient sample size.
- There were concerns whether the results of the correlation analysis performed to demonstrate score-level validity were correctly calculated. These correlations were used to demonstrate construct validity between the composite measure and its components and the composite and other similar immunization measures.

Measures Not Discussed by the Subgroup

2377 Defect Free Care for AMI

Scientific Methods Panel Votes: Measure passes

- Reliability: H-4, M-0, L-1, I-0
- Validity: H-3, M-1, L-1, I-0
- Composite H-2, M -2, L -1, I -0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below.

Reliability

- Reliability testing was conducted at both the data element and measure score levels.
- Testing of the data elements
  - Accomplished via the registry's audit program
  - The developer conducted an inter-rater reliability analysis that compared the values in the registry with those obtained by a trained abstractor; sample size was 330
  - Results: Reported kappa values ranged from 0.384 to 0.987 (aspirin in first 24 hours=0.384; cardiac rehab referral=0.386)
- Testing of the measure score
  - The developer used a split-sample methodology with data from 2016-17.
  - Results: Pearson correlation coefficient= 0.97.

Validity

- Empirical validity testing was conducted at the measure score level. Developers also described the conduct of a face validity assessment; however, that assessment does not conform to NQF's requirements.
- Developers conducted a construct validation analysis by correlating the results of this measure with results from a 30-day AMI mortality measure using Q4 2013-Q3 2014 data.

    - Developers hypothesized that use of defect free care processes for AMI patients leads to better outcomes.
    - Results: Pearson correlation coefficient = -0.1093 (statistically significant)
        - These results supported the developers' hypothesis, although the magnitude of the correlation was low.
        - The developer suggests that the low correlation value may be due to comparing a process measure to outcome measures or to other unmeasured factors that could contribute to the mortality results.
- Additional concerns regarding validity:
    - One Methods Panel member noted the two different target populations (STEMI and NSTEMI) and suggested that facility performance can be affected by the relative frequency of STEMI and NSTEMI.
    - The Panel noted that while developers presented data showing wide variation in performance across facilities, they did not test for statistically significant differences.
    - Panel members noted that hospitals that do not pass the data quality review for the NCDR registry are not included in the measure.

Composite

- Developers computed hospital-level results for the various components and correlated them with the composite results (via the Pearson correlation statistic). They found mostly moderate correlations (values ranged from 0.12 to 0.94).
- Panel members expressed some concern that the correlations were difficult to interpret, particularly given the different target populations. They would have liked to have seen a sensitivity analysis (e.g., to assess weighing scheme and inclusion of components).

2459 In-hospital Risk Adjusted Rate of Bleeding Events for patients undergoing PCI

Scientific Methods Panel Votes: Measure passes Reliability and Validity

- Reliability: H-4 , M-1, L-0, I-0
- Validity: H-4 , M-0, L-1, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below.

Reliability

- Reliability testing was conducted at both the data element and measure score levels.
- Data Element testing was conducted for some, but not all, critical data elements
    - Developers conducted a test-retest analysis by reviewing data for CathPCI patients who were readmitted or had a repeat procedure in 2016 (n=42,637). They analyzed 7 data for which values, in general, were not expected to change over the relatively short timeframe (i.e., elements (gender, age, cerebrovascular disease, peripheral vascular disease, chronic lunch disease, prior PCI, and diabetes).

    - Results: Inconsistencies in values for the 7 data elements ranged from 0.06% to 3%.
- Score-level testing was conducted using a signal-to-noise (SNR) analysis (specifically, Adams' beta-binomial method)
    - Results: Developers presented reliability estimates (presumably averages), for all procedures, by hospital volume terciles, and for hospitals with greater than average volume.
        - Values ranged from .706 to .819.
        - Panel members would have liked to see information about the variation in reliability estimates as well.

Validity

- Empirical validity testing was conducted at the measure score level.
    - NOTE: Developers also described face validity assessments through various means. It is possible that at least one of the assessments described when the measure was initially endorsed conforms to NQF requirements for face validity. However, those results were not presented and therefore were not considered when rating validity (moreover, face validity assessments are less important when results of empirical testing are available).
- Testing of the measure score
    - Developers conducted a construct validation analysis by examining the association of this measure (by quintile) with other outcome including mortality, complications of heart failure and stroke, length of stay, and rates of same-day discharge.
        - Developers hypothesized that that hospitals with higher bleeding rates would have higher rates on these other adverse outcomes measures.
        - Results: Developers found statistically significant associations between quintiles of bleeding rates and the outcomes of interest (higher rates of bleeding were associated with poorer outcomes). These results support the developers' hypothesis.
- This measure is risk-adjusted using hierarchical logistic regression with 32 risk factors.
    - Developers provided a conceptual rationale regarding why they did not include social risk factors in the risk-adjustment approach (i.e., the measure assesses in-hospital bleeding rate).
    - Some panel members questioned the lack inclusion of social risk factors in the risk-adjustment approach.
    - Model discrimination: C-statistic=0.79 for re-calibrated model using data from 2016 for 1,619 hospitals). (NOTE: c-statistic= 0.78 for initial model developed using data from 2/2008-4/2011 for 1,142 hospitals)
    - Model calibration: Developers assessed risk-model calibration by plotting observed versus predicted values. They report a slope=1 and intercept=0.

Subgroup 4: Initial Evaluation Call, Monday, October 15 from 2-4 pm ET; Follow-Up Evaluation Call, Tuesday, October 16 from 2-4 pm ET

During its initial call, Subgroup 4 discussed two measures (3452 and 3461) and accepted the preliminary analysis decisions for five measures (0167, 0174, 0175, 0176 and 0177) without further discussion. The subgroup discussed two additional measures (3227 and 3476) during the follow-up call. The final results of the nine measures evaluated by Subgroup 4 are presented below.

Measures Discussed by the Subgroup

3452 Access to Independence Promoting Services for Dual Eligible Beneficiaries

Scientific Methods Panel Votes: Measure does not pass

- Reliability: H-0, M-1, L-2, I-1
- Validity: H-0, M-1, L-2, I-1
- Composite: H-0, M-2; L-1, I-1

This measure was reviewed by the Methods Panel and did not pass. The Methods Panel raised concerns about the lack of data element reliability and validity testing, the results of score-level reliability testing, and the developer's risk adjustment approach.

3461 Functional Status Change for Patients with Neck Impairments

Scientific Methods Panel Votes: Measure did not pass on reliability and validity

- Reliability: H-0, M-3, L-2, I-0
- Validity: H-0, M-0, L-1, I-4

This measure was reviewed by the Methods Panel and did not pass. A summary is provided below:

Reliability

- Lack of specificity on data sources and definitions for key data elements
- Reliability testing methodology and results were unclear

Validity

- Lack of testing at the score level for both clinician and clinic level of analysis.
- Concerns regarding exclusions, the risk-adjustment methodology, handling of missing data, and meaningful differences.

3227 CollaboRATE Shared Decision Making Score

Scientific Methods Panel Votes: Measure does not pass on Reliability

- Reliability: H-0, M-1, L-1, I-3
- Validity: H-1, M-3, L-1, I-0

This measure was reviewed by NQF's Scientific Methods Panel. The panel agreed that the measure, as submitted, does not meet NQF's requirements for reliability due to concerns with the results from the score-level reliability testing.

3476 Communication Climate Assessment Toolkit

Scientific Methods Panel Votes: Measure does not pass

- Reliability: H-1, M-0, L-1, I-3
- Validity: H-2, M-2, L-0, I-1

This measure was reviewed by the Methods Panel. The panel agreed that the measure, as submitted, does not meet NQF's requirements for reliability due to lack of score-level reliability testing.

Measures not discussed by the subgroup

0167 Improvement in ambulation/locomotion

Scientific Methods Panel Votes: Measure passes Reliability and Validity

- Reliability: H-0, M-4, L-0, I-1
- Validity: H-1, M-4, L-0, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below.

Reliability

- Reliability testing was conducted at the both the data element and measure score levels.
- Testing of the data elements
    - Developers conducted an inter-rater reliability (IRR) analysis among nurses and physical therapists using a linear weighted kappa statistic. Testing of OASIS-C2 item M1860 was done using 2016-2017 data from home health patients in 4 states.
        - Start Of Care/Resumption Of Care: kappa=0.43 (n=105 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
        - Discharge: kappa=0.67 (n=83 patients) ["Substantial" agreement, according to the Landis and Koch classification system]
- Testing of the measure score
    - Developers used two approaches to assess reliability of the measure score: a signal-to-noise analysis using the Adams beta-binomial method and a split-sample analysis using ICC(2,1) and ICC(3,1) statistics. CY2016 data were used in testing.
        - Signal-to-noise reliability estimates: Mean=0.91; minimum=0.61; 10^{th} percentile = 0.77; median =0.95; 90^{th} percentile =1.00
        - Split sample reliability estimates: IRR(2,1)= 0.865; IRR(3,1)= 0.865 [NOTE that testing data limited to agencies with ≥40 qualifying episodes]
        - Panel members would like to have seen data element validation for variables included in the risk-adjustment model (and any other critical data elements).

Validity

- Validity testing was conducted at the measure score level. The developer also described various data element validation assessments; however, results of these assessments were only summarized, not presented.
- Developers conducted a construct [convergent] validation analysis by correlating (using the Spearman's rank correlation coefficient) the results of this measure with 4 other OASIS performance measures (improvement in bathing, bed transfer, and pain interfering with activity, and management of oral medications) and a modified version of the Quality of Patient Care Star Rating measure (modified by excluding the ambulation/locomotion measure from the calculation).
    - Developers expected statistically significant, strong, positive correlations.
    - Correlations with the 4 OASIS measures ranged from 0.61-0.82.
    - Correlation with the modified star-rating measure = 0.72.
    - These results aligned with supported the developers' hypothesis.
- This measure is risk-adjusted using logistic regression with 120 risk factors (based on 2016 data).
    - Developers discussed previous research linking dual-eligibility status and rural location with use of home health services. They therefore conducted analyses to examine associations between payment source (as a proxy for dual-eligibility) and rurality with this measure. They do include payment source in the risk-adjustment approach, but not rurality.
    - Model discrimination:
        - Overall development sample: c-statistic=0.779
        - Overall model validation sample: c-statistic= 0.779
    - Developers assessed risk-model calibration by calculating McFadden's R^2 and developing risk-decile plots.
        - Overall development sample: McFadden's R^2=0.174
        - Overall model validation sample: McFadden's R^2=0.167
- Panel members expressed some concern with excluding transferred patients, questioning whether those patients might have poorer outcomes on this measure. They had a similar concern with excluding patients who died.

0174 Improvement in bathing

Scientific Methods Panel Votes: Measure passes Reliability and Validity

- Reliability: H-0, M-4, L-0, I-1
- Validity: H-1, M-4, L-0, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below.

Reliability

- Reliability testing was conducted at the both the data element and measure score levels.
- Testing of the data elements
    - Developers conducted an inter-rater reliability (IRR) analysis among nurses and physical therapists using a linear weighted kappa statistic. Testing of OASIS-C2

item M1830 was done using 2016-2017 data from home health patients in 4 states.
    - Start Of Care/Resumption Of Care: kappa=0.51 (n=104 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
    - Discharge: kappa=0.43 (n=83 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
- Testing of the measure score
    - Developers used two approaches to assess reliability of the measure score: a signal-to-noise analysis using the Adams beta-binomial method and a split-sample analysis using ICC(2,1) and ICC(3,1) statistics. CY2016 data were used in testing.
        - Signal-to-noise reliability estimates: Mean=0.93; minimum=0.64; 10^{th} percentile = 0.80; median =0.96; 90^{th} percentile =0.99
        - Split sample reliability estimates: IRR(2,1)= 0.89; IRR(3,1)= 0.89 [NOTE that testing data limited to agencies with ≥40 qualifying episodes]
        - Panel members would like to have seen data element validation for variables included in the risk-adjustment model (and any other critical data elements).

<u>Validity</u>

- Validity testing was conducted at the measure score level. The developer also described various data element validation assessments; however, results of these assessments were only summarized, not presented.
- Developers conducted a construct [convergent] validation analysis by correlating (using the Spearman's rank correlation coefficient) the results of this measure with 4 other OASIS performance measures (improvement in ambulation/locomotion, bed transfer, and pain interfering with activity, and management of oral medications) and a modified version of the Quality of Patient Care Star Rating measure (modified by excluding the bathing measure from the calculation).
    - Developers expected statistically significant, strong, positive correlations.
    - Correlations with the 4 OASIS measures ranged from 0.68-0.82.
    - Correlation with the modified star-rating measure = 0.76.
    - These results aligned with supported the developers' hypothesis.
- This measure is risk-adjusted using logistic regression with 120 risk factors (based on 2016 data).
    - Developers discussed previous research linking dual-eligibility status and rural location with use of home health services. They therefore conducted analyses to examine associations between payment source (as a proxy for dual-eligibility) and rurality with this measure. They do include payment source in the risk-adjustment approach, but not rurality.
    - Model discrimination:
        - Overall development sample: c-statistic=0.760
        - Overall model validation sample: c-statistic= 0.760
    - Developers assessed risk-model calibration by calculating McFadden's R^2 and developing risk-decile plots.

    - Overall development sample: McFadden's R^2=0.152
    - Overall model validation sample: McFadden's R^2=0.147
- Panel members expressed some concern with excluding transferred patients, questioning whether those patients might have poorer outcomes on this measure. They had a similar concern with excluding patients who died.

0175 Improvement in bed transferring

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0, M-4, L-0, I-1
- Validity: H-1, M-4, L-0, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below.

Reliability

- Reliability testing was conducted at the both the data element and measure score levels.
- Testing of the data elements
    - Developers conducted an inter-rater reliability (IRR) analysis among nurses and physical therapists using a linear weighted kappa statistic. Testing of OASIS-C2 item M1850 was done using 2016-2017 data from home health patients in 4 states.
        - Start Of Care/Resumption Of Care: kappa=0.42 (n=104 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
        - Discharge: kappa=0.45 (n=83 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
- Testing of the measure score
    - Developers used two approaches to assess reliability of the measure score: a signal-to-noise analysis using the Adams beta-binomial method and a split-sample analysis using ICC(2,1) and ICC(3,1) statistics. CY2016 data were used in testing.
        - Signal-to-noise reliability estimates: Mean=0.92; minimum=0.65; 10^{th} percentile = 0.80; median =0.96; 90^{th} percentile =0.99
        - Split sample reliability estimates: IRR(2,1)= 0.89; IRR(3,1)= 0.89 [NOTE that testing data limited to agencies with ≥40 qualifying episodes]
- Panel members would like to have seen data element validation for variables included in the risk-adjustment model (and any other critical data elements).

Validity

- Validity testing was conducted at the measure score level. The developer also described various data element validation assessments; however, results of these assessments were only summarized, not presented.

- Developers conducted a construct [convergent] validation analysis by correlating (using the Spearman's rank correlation coefficient) the results of this measure with 4 other OASIS performance measures (improvement in ambulation/locomotion, bathing, and pain interfering with activity, and management of oral medications) and a modified version of the Quality of Patient Care Star Rating measure (modified by excluding the bed transferring measure from the calculation).
    - Developers expected statistically significant, strong, positive correlations.
    - Correlations with the 4 OASIS measures ranged from 0.52-0.70.
    - Correlation with the modified star-rating measure = 0.65.
    - These results aligned with supported the developers' hypothesis.
- This measure is risk-adjusted using logistic regression with 113 risk factors (based on 2016 data).
    - Developers discussed previous research linking dual-eligibility status and rural location with use of home health services. They therefore conducted analyses to examine associations between payment source (as a proxy for dual-eligibility) and rurality with this measure. They do include payment source in the risk-adjustment approach, but not rurality.
    - Model discrimination:
        - Overall development sample: c-statistic=0.792
        - Overall model validation sample: c-statistic= 0.792
    - Developers assessed risk-model calibration by calculating McFadden's R^2 and developing risk-decile plots.
        - Overall development sample: McFadden's R^2=0.198
        - Overall model validation sample: McFadden's R^2=0.190
- Panel members expressed some concern with excluding transferred patients, questioning whether those patients might have poorer outcomes on this measure. They had a similar concern with excluding patients who died.

0176 Improvement in management of oral medications

Scientific Methods Panel Votes: Measure passes

- Reliability: H-3, M-1, L-0, I-1
- Validity: H-1, M-4, L-0, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below.

Reliability

- Reliability testing was conducted at the both the data element and measure score levels.
- Testing of the data elements
    - Developers conducted an inter-rater reliability (IRR) analysis among nurses and physical therapists using a linear weighted kappa statistic. Testing of OASIS-C2 item M2020 was done using 2016-2017 data from home health patients in 4 states.
        - Start Of Care/Resumption Of Care: kappa=0.59 (n=105 patients) ["Moderate" agreement, according to the Landis and Koch classification

system] [NOTE that submission form also reported a kappa of 1.0, but this was likely a typo.]
        - Discharge: kappa=0.65 (n=84 patients) ["Substantial" agreement, according to the Landis and Koch classification system]
- Testing of the measure score
    - Developers used two approaches to assess reliability of the measure score: a signal-to-noise analysis using the Adams beta-binomial method and a split-sample analysis using ICC(2,1) and ICC(3,1) statistics. CY2016 data were used in testing.
        - Signal-to-noise reliability estimates: Mean=0.92; minimum=0.68; 10^{th} percentile = 0.80; median =0.95; 90^{th} percentile =0.99
        - Split sample reliability estimates: IRR(2,1)= 0.89; IRR(3,1)= 0.89 [NOTE that testing data limited to agencies with ≥40 qualifying episodes]
- Panel members would like to have seen data element validation for variables included in the risk-adjustment model (and any other critical data elements).

<u>Validity</u>

- Validity testing was conducted at the measure score level. The developer also described various data element validation assessments; however, results of these assessments were only summarized, not presented.
- Developers conducted a construct [convergent] validation analysis by correlating (using the Spearman's rank correlation coefficient) the results of this measure with 4 other OASIS performance measures (improvement in ambulation/locomotion, bathing, bed transfer, and pain interfering with activity) and a modified version of the Quality of Patient Care Star Rating measure (modified by excluding the improvement of management of oral medications measure from the calculation).
    - Developers expected statistically significant, strong, positive correlations.
    - Correlations with the 4 OASIS measures ranged from 0.51-0.68.
    - Correlation with the modified star-rating measure = 0.62.
    - These results aligned with supported the developers' hypothesis.
- This measure is risk-adjusted using logistic regression with 117 risk factors (based on 2016 data).
    - Developers discussed previous research linking dual-eligibility status and rural location with use of home health services. They therefore conducted analyses to examine associations between payment source (as a proxy for dual-eligibility) and rurality with this measure. They do include payment source in the risk-adjustment approach, but not rurality.
    - Model discrimination:
        - Overall development sample: c-statistic=0.777
        - Overall model validation sample: c-statistic= 0.777
    - Developers assessed risk-model calibration by calculating McFadden's R^2 and developing risk-decile plots.
        - Overall development sample: McFadden's R^2=0.182
        - Overall model validation sample: McFadden's R^2=0.179

- Panel members expressed some concern with excluding transferred patients, questioning whether those patients might have poorer outcomes on this measure. They had a similar concern with excluding patients who died.

0177 Improvement in pain interfering with activity

Scientific Methods Panel Votes: Measure passes

- Reliability: H-1, M-3, L-0, I-1
- Validity: H-1, M-4, L-0, I-0

This measure was reviewed by the Scientific Methods Panel. A summary of the measure is provided below.

Reliability

- Reliability testing was conducted at the both the data element and measure score levels.
- Testing of the data elements
    - Developers conducted an inter-rater reliability (IRR) analysis among nurses and physical therapists using a linear weighted kappa statistic. Testing of OASIS-C2 item M1242 was done using 2016-2017 data from home health patients in 4 states.
        - Start Of Care/Resumption Of Care: kappa=0.45 (n=105 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
        - Discharge: kappa=0.53 (n=84 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
- Testing of the measure score
    - Developers used two approaches to assess reliability of the measure score: a signal-to-noise analysis using the Adams beta-binomial method and a split-sample analysis using ICC(2,1) and ICC(3,1) statistics. CY2016 data were used in testing.
        - Signal-to-noise reliability estimates: Mean=0.95; minimum=0.74; 10^{th} percentile = 0.87; median =0.97; 90^{th} percentile =1.00
        - Split sample reliability estimates: IRR(2,1)= 0.90; IRR(3,1)= 0.90 [NOTE that testing data limited to agencies with ≥40 qualifying episodes]
- Panel members would like to have seen data element validation for variables included in the risk-adjustment model (and any other critical data elements).

Validity

- Validity testing was conducted at the measure score level. The developer also described various data element validation assessments; however, results of these assessments were only summarized, not presented.

- Developers conducted a construct [convergent] validation analysis by correlating (using the Spearman's rank correlation coefficient) the results of this measure with 4 other OASIS performance measures (improvement in ambulation/locomotion, bathing, and bed transfer, and management of oral medications) and the Quality of Patient Care Star Rating measure [it is unclear whether this was a modified version of the measure that excluded the pain measure]
    - Developers expected statistically significant, strong, positive correlations.
    - Correlations with the 4 OASIS measures ranged from 0.51-0.69.
    - Correlation with the star-rating measure = 0.65.
    - These results aligned with supported the developers' hypothesis.
- This measure is risk-adjusted using logistic regression with 114 risk factors (based on 2016 data).
    - Developers discussed previous research linking dual-eligibility status and rural location with use of home health services. They therefore conducted analyses to examine associations between payment source (as a proxy for dual-eligibility) and rurality with this measure. They do include payment source in the risk-adjustment approach, but not rurality.
    - Model discrimination:
        - Overall development sample: c-statistic=0.656
        - Overall model validation sample: c-statistic= 0.657
    - Developers assessed risk-model calibration by calculating McFadden's R^2 and developing risk-decile plots.
        - Overall development sample: McFadden's R^2=0.053
        - Overall model validation sample: McFadden's R^2=0. 051
- Panel members expressed some concern with excluding transferred patients, questioning whether those patients might have poorer outcomes on this measure. They had a similar concern with excluding patients who died.