# MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

**To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return**

## Brief Measure Information

**NQF #:** 1789
**Corresponding Measures:**
De.2. **Measure Title:** Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)
Co.1.1. **Measure Steward:** Centers for Medicare & Medicaid Services (CMS)
De.3. **Brief Description of Measure:** For the hospital-wide readmission (HWR) measure that was previously endorsed and is used in the Hospital Inpatient Quality Reporting Program (IQR), the measure estimates a hospital-level risk-standardized readmission rate (RSRR) of unplanned, all-cause readmission after admission for any eligible condition within 30 days of hospital discharge. The measure reports a single summary RSRR, derived from the volume-weighted results of five different models, one for each of the following specialty cohorts based on groups of discharge condition categories or procedure categories: surgery/gynecology; general medicine; cardiorespiratory; cardiovascular; and neurology, each of which will be described in greater detail below. The measure also indicates the hospital-level standardized risk ratios (SRR) for each of these five specialty cohorts. The outcome is defined as unplanned readmission for any cause within 30 days of the discharge date for the index admission (the admission included in the measure cohort). A specified set of planned readmissions do not count in the readmission outcome. CMS annually reports the measure for patients who are 65 years or older, are enrolled in fee-for-service (FFS) Medicare, and hospitalized in non-federal hospitals.

For the All-Cause Readmission (ACR) measure version used in the Shared Savings Program (SSP), the measure estimates an Accountable Care Organization (ACO) facility-level RSRR of unplanned, all-cause readmission after admission for any eligible condition within 30 days of hospital discharge. The ACR measure is calculated using the same five specialty cohorts and estimates an ACO-level standardized risk ratio for each. CMS annually reports the measure for patients who are 65 years or older, are enrolled in FFS Medicare and are ACO assigned beneficiaries.
1b.1. **Developer Rationale:** The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, ACOs, and policy makers with information about risk-standardized all-cause unplanned readmission rates among Medicare beneficiaries 65 years and older admitted to all non-federal US acute care hospitals. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions' whose performance is better or worse than would be expected based on their patient case mix and hospital service mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

Hospital-wide readmission is a priority area for outcomes measure development as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting readmission

rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices, as well as increase transparency for consumers.

For the ACR measure, several ACOs have shared with CMS the interventions they have implemented to reduce hospital readmissions. ACOs are redesigning care to improve results on the ACR measure. Some specific examples include:

1. Care coordination focusing on transitions or special populations
One ACO works to prevent readmissions to the hospital through the Transitions of Care program. A medical assistant care transition navigator conducts telephone outreach to patients at 48 hours and two weeks post-discharge.

Another ACO focuses on reducing readmissions via a home connection program for high-risk populations. Key components include ensuring a physician follow-up appointment is scheduled before discharge, ensuring patients have a personal contact for urgent needs, and ensuring patients understand how to manage their medications. Many ACOs focus on improved transitions of care for patients with end-stage renal disease to prevent readmissions.

2. Pharmacy involvement: Strategies include medication reconciliation as well as data integration with labs and pharmacies. Another strategy is increased pharmacist involvement in transitions of care: One ACO has a pharmacist focusing on transitions of care to reduce readmissions for patients with heart failure, Chronic Obstructive Pulmonary Disease (COPD), and pneumonia.

S.4. **Numerator Statement:** The outcome for the HWR measure is 30-day readmission. We define readmission as an inpatient admission for any cause, with the exception of certain planned readmissions, within 30 days from the date of discharge from an eligible index admission. If a patient has more than one unplanned admission (for any reason) within 30 days after discharge from the index admission, only one is counted as a readmission. The measure looks for a dichotomous yes or no outcome of whether each admitted patient has an unplanned readmission within 30 days. However, if the first readmission after discharge is considered planned, any subsequent unplanned readmission is not counted as an outcome for that index admission because the unplanned readmission could be related to care provided during the intervening planned readmission rather than during the index admission.

The outcome for the ACR measure is also 30-day readmission. The outcome is defined identically to what is described above for the HWR measure.

S.7. **Denominator Statement:** The measure at the hospital level includes admissions for Medicare beneficiaries who are 65 years and older and are discharged from all non-federal, acute care inpatient US hospitals (including territories) with a complete claims history for the 12 months prior to admission.

The measure at the ACO level includes all relevant admissions for ACO assigned beneficiaries who are 65 and older and are discharged from all non-Federal short-stay acute care hospitals, including critical access hospitals.

Additional details are provided in S.9 Denominator Details.

S.10. **Denominator Exclusions:** The measure excludes index admissions for patients:

1. Admitted to Prospective Payment System (PPS)-exempt cancer hospitals;
2. Without at least 30 days post-discharge enrollment in FFS Medicare;
3. Discharged against medical advice (AMA);
4. Admitted for primary psychiatric diagnoses;
5. Admitted for rehabilitation; or

| |
|---|
| 6. Admitted for medical treatment of cancer. |
| De.1. **Measure Type:** Outcome<br>S.23. **Data Source:** Claims (Only)<br>S.26. **Level of Analysis:** Facility, Integrated Delivery System |
| **IF Endorsement Maintenance – Original Endorsement Date:** Apr 24, 2012 **Most Recent Endorsement Date:** Dec 09, 2016 |
| **IF this measure is included in a composite, NQF Composite#/title:**<br><br>**IF this measure is paired/grouped, NQF#/title:**<br><br>De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A |

## Preliminary Analysis

| |
|---|
| To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion. |
| **Criteria 1: Importance to Measure and Report** |
| **1a. Evidence**<br>**Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.** |
| **1a. Evidence.** The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.<br><br>• The developer suggests that healthcare providers are able to influence readmission rates through a broad range of clinical activities including communication between providers, prevention of, and response to, complications, patient safety and coordinated transitions to the outpatient environment.<br>• The developer provides specific examples of successful interventions accountable care organizations (ACOs) have implemented to reduce hospital readmission rates including:<br>  ○ Care coordination focusing on transitions<br>  ○ Home connection programs for high-risk populations<br>  ○ Working with pharmacies to ensure medication reconciliation as well as data integration.<br><br>Guidance **from the Evidence Algorithm**<br>Evidence Algorithm guidance: 1) Measure assesses a health outcome → 2) The relationship between the health outcome and at least one healthcare action is identified and supported by the stated rationale →PASS<br><br>***Question for the Committee:*** |

○ *Is there at least one clinical action that the provider can undertake to achieve a change in the measure results?*

**Preliminary rating for evidence:** ☒ **Pass** ☐ **No Pass**

**1b. Gap in Care/Opportunity for Improvement and 1b. Disparities**
**Maintenance measures – increased emphasis on gap and variation**

**1b. Performance Gap.** The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer provides performance data for this ACO measure from three calendar years (2013, 2014, and 2015), covering approximately 3,900,000 admissions.
- The data show that during the calendar year of 2015, readmission rates ranged from a minimum of 13.1% to a maximum of 17.5%, with the 10$^{th}$ percentile at 14.0%, the 50$^{th}$ percentile at 14.8%, and the 90$^{th}$ percentile at 15.7%.

**Disparities**

- To help in assessment of potential disparities, the developers also provided performance scores (using 2015 data) for ACOs serving a low proportion of dual eligible patients vs. those serving a high proportion of dual eligible patients and performance scores for ACOs serving a low proportion of patients below median AHRQ SES Index Score vs. those serving a high proportion of patients below the median AHRQ SES index score.

- ACOs serving a low proportion (=5.1%) Dual Eligible patients had a slightly lower median readmission rates (-0.7%) compared to ACOs serving a high proportion (=13.3%) Dual Eligible patients.

- ACOs serving a low proportion of patients below median AHRQ SES index score had slightly lower median readmissions rates (-0.1%) compared to ACOs serving a high proportion of patients below median AHRQ SRS index score.

- By proportion of **Dual Eligible Patients:**

Distribution of RSRRs for the ACR measure by Proportion of Dual-Eligible Patients
Dates of Data: January 1, 2015- December 31, 2015
Data Source: Medicare FFS claims for ACO assigned/aligned beneficiaries.
Characteristic//ACOs with a low proportion (=5.1%) Dual Eligible patients// ACOs with a high proportion (=13.3%) Dual Eligible patients
Number of ACOs// 103// 104
Number of Patients// 247,252 in low-proportion of ACOs // 217,145 in high-proportion ACOs
Maximum// 16.2 // 17.5
90th percentile// 15.4 // 16.2
75th percentile// 14.9 // 15.7
Median (50th percentile)// 14.6 // 15.3
25th percentile// 14.2 // 14.7
10th percentile// 13.9 // 14.5
Minimum// 13.1 // 13.8

- By proportion of **below median AHRQ SES index score:**

Characteristic// ACOs with a low proportion of patients below median AHRQ SES index score (=31.8%)// ACOs with a high proportion of patients below median AHRQ SES index score (=66.4%)
Number of ACOs// 104 // 104
Number of patients// 336,504 // 185,644
Maximum// 17.2 // 16.7
90th percentile// 15.9 // 15.9
75th percentile// 15.4 // 15.4
Median// 14.8 // 14.9
25th percentile// 14.4 // 14.6
10th percentile// 14.1 // 14.2
Minimum // 13.1 // 13.5

***Questions for the Committee:***

- *Is there a sufficient performance gap that warrants a national performance measure in this topic area for ACOs?*

**Preliminary rating for opportunity for improvement: ☐ High ☒ Moderate ☐ Low ☐ Insufficient**

## Committee pre-evaluation comments

**Criteria 1: Importance to Measure and Report (including 1a and 1b)**

***1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission?For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.***

Comments:

** Healthcare providers should be able to influence readmission rates through a range of clinical activities such as care coordination, home connection programs for high risk populations and data integration.

** Acceptable

** The developers present sound evidence in support of the HWRs conceptual import, but could do more to support its construct, particularly in terms of included covariates, and whether the limited between hospital/ACO variation might be explained away with closer to perfect information (non-claims based, SDOH, etc.). In addition, the evidence attachment is dated 1/2016. In 10/2017, Zuckerman and colleagues found that moving the HRRP assessments to the HWR measure would result in significantly higher penalties for safety-net hospitals (N Engl J Med 2017; 377:1551-1558). Is the same true for safety-net equivalent ACOs under the ACR measure?

** Given that this is not a new measure, there is/has been consistent evidence that outcome measured can be improved by health system interventions.

** The evidence is dependent on access and acceptance of post acute services that may not be in control of the hospital. Dual eligible patients or those patients with co-morbid conditions of mental health, SUD, or Z codes are less likely to be affected by simply transition planning

** For the ACR measure for ACO facility level RSRRs, the developers proved examples of efforts from interventions from CMS demonstration projects to reduce readmissions, including care coordination at transition, and pharmacy involvement in medication reconciliation and transitions of care. The developers also provided a conceptual model linking specific strategies for reducing readmissions to decrease in risk of readmission (see Figure 1a.2.).

***1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?***

Comments:

** Data shows a range of readmission rates from a minimum of 13.1% to maximum of 17.5%. Percentile ranges at the 10th percentile at 14%, 50th percentile at 14.8 percentile and 90th percentile at 15.7%. ACOs with low proportion of dual eligible patients had a slightly lower median readmission rate as compared to ACOs serving a high proportion Dual Eligible patients.

** Gap acceptable Disparities exist- reporting should continue to include stratification by SES variables

** More information is needed from the developers on their treatment of disparities. In response to 1b.4 (required for endorsement), little information is provided on categorical groupings of low vs. high proportion dual hospitals and ACOs (are these top and bottom quartiles, and is linearity a sound assumption if so?). The importance of a thoughtful evaluation of the role of SDOH in the ACR measure is magnified because 1) the measure is used to allocate dollars, and 2) the measure does not fall under the establishing beneficiary equity component of the 21st Century Cures Act. As one technical committee reviewer stated "The developers state that social factors do not make much difference in the models, but their actual results suggest otherwise....Since the distribution for this measure is so tight, that suggests that half of the ACOs would move across decile scoring categories with adjustment, and at least one would move from the bottom to top. I would suspect the ACOs thus affected would argue that adjustment would make a meaningful difference". Disincentivizing ACO entrants (or incentivizing departures) in markets and communities with social disadvantage is something I think all would agree to be a very undesirable, but real risk. The issue deserves more diligence than provided.

** There remains a performance gap, and data were presented. It appears that the differences in performance have tightened up. Agree with one preliminary review that there does seem to be some difference in systems that serve a higher proportion of duals, which might speak to disparities in care.

** Spread might be expected and explained by disparities

** As noted in the reviews by the Scientific Methods Panel, the distribution of scores is very narrow. However, as noted in the response by the developers, a one-unit change in RSRR translates to 5,000 beneficiaries which to them is a clinically meaningful drop. They also not that the maximum improvement under the Pioneer and CMS SSP ACO programs was 12.3%, with a mean of 3/4% from 2014-2015. Data on disparities is largely provided for the HWR measure. Disparities in ACRs for ACOs were noted for dual-eligibles, and for the AHRQ SES indicator. However, adjustment for these variables produced very little change (<.11%) in ACO's RSRRs.

**Criteria 2: Scientific Acceptability of Measure Properties**

**2a. Reliability: Specifications and Testing**
**2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data**

**Reliability**
**2a1. Specifications** requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.
**2a2. Reliability testing** demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.
**Validity**
**2b2. Validity testing** should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.
**2b2-2b6. Potential threats to validity** should be assessed/addressed.
**Composite measures only:**
**2d. Empirical analysis to support composite construction**. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

**Complex measure evaluated by Scientific Methods Panel**? ☒ **Yes** ☐ **No**
**Evaluators:** Karen Joynt Maddox, Jennifer Perloff, Jack Needleman, David Cella, Bijan Borah

**Evaluation of Reliability and Validity (and composite construction, if applicable)**: Link A, Link B, Link C, Link D, Link E

***Questions for the Committee regarding reliability:***
- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

***Questions for the Committee regarding validity:***
- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

**Preliminary rating for reliability:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

**Preliminary rating for validity:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 2: Scientific Acceptability of Measure Properties (including all 2a and 2b)**

***2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?***

Comments:

** Reliability is moderate. Noted difference in reliability between the ACO and the hospital versions of the ACR measure. ACO reliability may be lower as ACOs typically have beneficiaries admitted to more than one hospital. ACOs. represent a higher-level organization often including admission from several hospitals. IC was 0.62-high reliability.

** low to moderate

** No issues re: ability to implement this measure, measure specs remain clear, though ?why dementia remains excluded by CMS

** Linking specific hospital related activities to a reliable reduction in readmission has not been proven. A bundle approach and access to services recommended in a gold standard transition plan is needed. Patient acceptance of the plan is also required

** The developers' strategy closely parallels that for the HWR measure. They provide great detail on measure description, coding, analytic algorithm, data sources, and sampling, such that as for the HWR measure, the ACR measure could be consistently implemented.

***2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?***

Comments:

** see previous question

** low to moderate

** no

** yes, above

** The ICC noted for the ACR measure were somewhat lower than that for the HWR measure, however, that is probably due to the error introduced by the 'higher order' nesting of patients within hospitals and subsequently hospitals with ACOs. It is rather surprising that the ICCs for the ACR are as high as the developers report. It would be useful to have confidence intervals around the ICCs.

***2b1. Validity -Testing: Do you have any concerns with the testing results? 2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?***

Comments:

** Moderate validity. Developers suggest that the social factors do not make much difference in the models, however actual results have a tight distribution.

** low moderate

** Risk should include other SES factors, perhaps Z-codes or co-morbid conditions that hinder the patients' ability to follow the care plan

** The developers assessed predictive validity, (comparing ACOs' ACRs with overall and care coordination/patient safety performance scores, RSRRs for diabetes, heart failure and chronic conditions in subsequent performance years), convergent validity (comparing ACRs with care coordination/patient safety measures in the same year), and discriminant validity (comparing ACRs with preventive care measures). Observed correlations were in the expected direction and were in the upper range of moderate (sharing roughly 25-30% variation) for predictive and concurrent validity, and small but significant for the prevention measures (sharing less than 5% variance). Missing data does not appear to constitute a threat to validity of the ACR measure.

***2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?***

Comments:

** Suggest to use the risk adjustment

** acceptable

** Standard exclusions--would be good to obtain group consensus about this. Still some degree of uncertainty re: the risk adjustment and whether or not social risk factors such as dual status should be included, esp as ACOs move increasingly into the Medicaid space

** I don't think dual eligibility captures all the social risks

** As for the HWR measure, although the risk prediction models show statistically significant relationships between the AHRQ SES index and dual eligibility with ACOs' ACR, adjustment for these variables did not substantially or substantively change the scores on this measure.

**Criterion 3. Feasibility**
**Maintenance measures – no change in emphasis – implementation issues may be more prominent**

**3. Feasibility** is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is based on administrative claims data (e.g., DRG, ICD-9/10), which the developers note are routinely generated and collected as part of hospitals' billing processes.
- The developer indicates that all data elements are in defined fields in electronic claims.

***Questions for the Committee:***

- *Are the required data elements routinely generated and used during care delivery?*

**Preliminary rating for feasibility:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 3: Feasibility**

***3. Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?***

Comments:

** From claims data

** feasible

** no concerns. This is not a new measure

** Yes

** None

**Criterion 4: Usability and Use**

**Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences**

**4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)**

**4a. Use** evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

**4a.1. Accountability and Transparency.** Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

**Current uses of the measure**

**Publicly reported?** ☒ **Yes** ☒ **No**

**Current use in an accountability program?** ☒ **Yes** ☐ **No**

**OR**

**Planned use in an accountability program?** ☐ **Yes** ☐ **No**

**Accountability program details**

The measure is used in the Medicare Shared Savings Program, the Pioneer ACO Mode, and the Next Generation ACO Model.

Medicare SSP: The Medicare SSP was established by Section 3022 of the Affordable Care Act. The SSP is a new approach to the delivery of health care. Through ACOs, the SSP facilitates coordination and cooperation among providers to improve the quality of care for Medicare FFS beneficiaries and lower the growth in Medicare expenditures. Eligible providers, hospitals, and suppliers may participate in the SSP by creating or participating in an ACO. The mean performance rate for SSP ACOs was 14.86 (range: 13.1-17.49) in 2015. As of January 2017, there are 480 SSP ACOs with over 9 million assigned beneficiaries across the 50 states, Puerto Rico, and Washington DC. An ACO may serve patients across multiple regions. ACOs include networks of individual practices, group practices, hospital/professional partnerships, hospitals employing ACO professionals, federally qualified health centers, rural health clinics, and critical access hospitals. An ACO may report multiple of these characteristics.

Pioneer ACO Model: The Pioneer ACO Model is designed for health care organizations and providers that are already experienced in coordinating care for patients across care settings. It will allow these provider groups to move more rapidly from a shared savings payment model to a population-based payment model on a track consistent with, but separate from, the Medicare SSP. It is designed to work in coordination with private payers by aligning provider incentives, which will improve quality and health outcomes for patients across the ACO, and achieve cost savings for Medicare, employers, and patients.

The mean performance rate for Pioneer ACOs was 15.41 (range: 13.98-16.71) in 2015. The Pioneer ACO Model began with 32 ACOs in 2012 and concluded December 31, 2016 with 8 ACOs participating. Pioneer ACOs are located across the US.
https://innovation.cms.gov/initiatives/Pioneer-aco-model/
Next Generation ACO Model: The Next Generation ACO Model is an initiative for ACOs that are experienced in coordinating care for populations of patients. It will allow these provider groups to assume higher levels of financial risk and reward than are available under the current Pioneer Model and SSP. The goal of the Model is to test whether strong financial incentives for ACOs, coupled with tools to support better patient engagement and care management, can improve health outcomes and lower expenditures for Original Medicare fee-for-service (FFS) beneficiaries.
https://innovation.cms.gov/initiatives/Next-Generation-ACO-Model/

Eighteen ACOs participated in the Next Generation ACO Model for the 2016 performance year, and twenty-eight ACOs are joining the Model for 2017.

ACOs voluntarily participate in the SSP/Pioneer ACO Model/Next Generation ACO Model following an application and CMS approval process. In these ACOs, the ACR measure reflects the RSRR at an ACO-level rather than a hospital level. ACOs vary substantially in composition and typically include multiple types of care delivery entities. For example, some ACOs are networks of group practices, some involve partnerships between ACO professionals and hospitals, and some are hospital-based ACOs. While ACOs may include hospitals as participants, ACOs are not required to have hospitals as participants. For

instance, as of January 2017, only 38% of ACOs participating in the SSP involve partnerships between ACO professionals and hospitals.

In the SSP/Pioneer ACO Model/Next Generation ACO Model, the ACR measure is pay for reporting 2 years before phasing into pay for performance. ACOs receive full points for pay for reporting measures when they completely report data to CMS. For the ACR measure, no quality reporting is required by ACOs because CMS uses administrative claims to calculate the measure. The pay for reporting period of 2 performance years gives ACOs the opportunity to become familiar with the measure and understand their performance before CMS begins assessing ACO performance against a measure benchmark. Under pay for performance, ACO performance is compared to a benchmark that is calculated using all Medicare FFS data. In addition, benchmarks are set for 2 performance years to give ACOs a steady target for improving performance. ACOs may also receive quality improvement points in calculating the ACO Overall Quality Score, if they demonstrate a significant improvement in performance from one year to the next. The ACR measure performance is included in these quality improvement point calculations.

**4a.2. Feedback on the measure by those being measured or others.** Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

**Feedback on the measure by those being measured or others:**

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *How has the measure been vetted in real-world settings by those being measured or others?*

**Preliminary rating for Use:** ☒ **Pass** ☐ **No Pass**

**4b. Usability (4a1. Improvement; 4a2. Benefits of measure)**

**4b. Usability** evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

**4b.1 Improvement.** Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

**Improvement results**

The developer reports that from 2012 to 2015, the ACO median RSRR decreased from 14.85 in 2012 to 14.83 in 2015. The developer noted that the number of ACOs participating in the SSP increased from 114 in 2012 to 397 ACOs in 2015. In 2015, 69.38% of Pioneer and SSP ACOs with two years of performance data improved their ACR rates. The maximum percent improvement for Pioneer and SSP ACOs was 12.3% with a mean of 3.4% from 2014 to 2015. Of those that improved, 33 SSP ACOs and 1

Pioneer Model ACO saw a significant enough improvement to receive additional points under the Quality Improvement Reward.

**4b2. Benefits vs. harms.** Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

**Unexpected findings (positive or negative) during implementation**

- The developer noted that there are no unexpected findings to report.

**Potential harms**

- The developer noted that there were no unintended consequences during development, testing or re-specification. They are committed to ongoing monitoring of potential unintended consequences over time.

**Feedback:**

- During the 2011-2012 MAP review, MAP recommended this measure be submitted for NQF review and endorsement.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

**Preliminary rating for Usability and use:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 4: Usability and Use**

***4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?***

Comments:

** Measure is sued in other programs such as the Medicare shared savings program, the Pioneer ACO mode and the next generation ACO model.

** acceptable

** These measures are already being used across many levels of care. At the local level, ACOs are already reporting on this measure.

** Limited

** Unclear.

***4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.***

Comments:

** Developers showed that the ACO median RRR decreased from 14.85 in 2012 to 14.83 in 2015. The number of ACOs participating in the SSP increased from 114 in 2012 to 397 ACOs in 2015.

** moderate concerns about unintended consequences and the impact of factors outside of the traditional scope of a hospital or hospital system, though perhaps less so for ACO

** benefits vs harms: we are not capturing what happens re: those patients who were never admitted, but who received 'hospital level' care in in an outpatient setting--do think some systems have gamed this measure to some extent by "obs-ing" pts rather than formally admitting them. Obs'd patients may incur higher personal costs, but overall for the hospitals and ACOs, it appears that costs were reduced as well as RE-admissions given that these patients were never admitted in the first place. However, since CMS started turning a lens on re-admissions, there has been a noticeable benefit in terms of additional supports for systems/patients to prevent unplanned readmissions (e.g. improved coordination and care management post-discharge) as well as a laudable increase in programs that seek to provide appropriate hospital level of care at home.

** Limited - unintended consequence could be penalties affecting those ACOs caring for safety net patients

** As with the HWR measure, there is the potential for disincentives for enrolling the poor, underserved beneficiaries and those from racial/ethnic minorities. Those ACOs serving those groups may be at greater risk for higher ACRs for reasons not attributable to quality of care provided but to characteristics of the patient and their environment beyond the control of the ACO.

**Criterion 5: Related and Competing Measures**

**Related or competing measures**

- 1768 : Plan All-Cause Readmissions (PCR)

**Harmonization**

- This measure and the NCQA Plan All-Cause Readmissions (PCR) Measure #1768 are related measures, but are not competing because they don't have the same measure focus and same target population. Each of these measures has different specifications. In addition, both have been previously harmonized to the extent possible under the guidance of the National Quality Forum Steering Committee in 2011.

# Measure Number: 1789
# Measure Title: Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

**Instructions for filling out this form:**

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

## RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   **REFERENCE:** "MIF_xxxx" document

   ***NOTE****: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

***TIPS**: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

**REFERENCE:** "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

**REFERENCE**: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

4. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   **REFERENCE:** Testing attachment, section 2a2.

   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☐Yes (go to Question #7)

   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   **REFERENCE:** Testing attachment, section 2a2.2

   ***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

   *Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

   ☐Yes (go to Question #8)

   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

   **REFERENCE:** Testing attachment, section 2a2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?

   **REFERENCE:** testing attachment section 2b1.

   **NOTE:** Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

   ***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*

   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

## OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

    ☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

    ☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

# VALIDITY

## Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

    **REFERENCE:** Testing attachment, section 2b2-2b6

    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

    ☒Yes (go to Question #12)

    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

12. Analysis of potential threats to validity: Any concerns with measure exclusions?

    **REFERENCE:** Testing attachment, section 2b2.

    ***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

    ☐Yes (please explain below then go to Question #13)

    ☒No (go to Question #13)

    ☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

    **REFERENCE:** Testing attachment, section 2b3.

    13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

    13b. Are social risk factors included in risk model? ☐Yes ☒No

    13c. Any concerns regarding the risk-adjustment approach?

    ***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*

*Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☒Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

The developers state that social factors do not make much difference in the models, but their actual results suggest otherwise, particularly given the very tight distribution of performance for this measure. For example, after adjusting for dual status, the median change was -0.105% (interquartile range [IQR] -0.305% – 0.216%, minimum -0.794% – maximum 2.426%). Since the distribution for this measure is so tight, that suggests that half of the ACOs would move across decile scoring categories with adjustment, and at least one would move from the bottom to top. I would suspect the ACOs thus affected would argue that adjustment would make a meaningful difference.

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
**REFERENCE:** Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

☐No (go to Question #15)

As above, scores are really tightly distributed, so depending how the measure is used, statistically significant differences may be clinically meaningless, particularly in a measure with relatively poor discriminative capability at baseline.

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
**REFERENCE:** Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
**REFERENCE:** Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

## Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

    **REFERENCE:** Testing attachment, section 2b1.

    ***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

    ☒Yes (go to Question #18)

    ☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?

    **REFERENCE:** Testing attachment, section 2b1.

    ***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

    ☒Yes (go to Question #19)

    ☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    **REFERENCE:** Testing attachment, section 2b1.

    ***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

    ☒Yes (go to Question #20)

    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #21)

    ☒Moderate (go to Question #21)

    ☐Low (please explain below then go to Question #21)

    ☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?

    **REFERENCE:** Testing attachment, section 2b1.

    ***TIPS**: Prior validity studies of the same data elements may be submitted*

    ☐Yes (go to Question #22)

    ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:*** *Testing attachment, section 2b1.*

***TIPS****: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

*Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

**NOTE:** If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

**REFERENCE:** Testing attachment, section 2b1.

***TIPS****: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS****: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

### OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of <u>all</u> testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there <u>are</u> threats to validity and/or threats to validity were <u>not assessed</u>]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level <u>is required</u>; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

## FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

**REFERENCE:** Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

# Measure Number: 1789
# Measure Title: Hospital-Wide Readmission Measure (HWR)

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

**Instructions for filling out this form:**

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

## RELIABILITY

28. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

**REFERENCE:** "MIF_xxxx" document

***NOTE****: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

***TIPS****: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

29. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

**REFERENCE:** "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

30. Was reliability testing conducted with computed performance measure scores for each measured entity?

**REFERENCE**: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

31. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

Test/re-test with a random sample of patients from the last year of data; ICC to compare results

32. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

The reliability is lower for the ACO measure than the hospital measure – this raises the question of why?

33. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

**REFERENCE:** Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

34. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

*Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☒Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

35. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

**REFERENCE:** Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

36. Was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

**REFERENCE:** testing attachment section 2b1.

**NOTE:** Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

***TIP:** You should answer this question <u>ONLY</u> if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.***

☐Yes (go to Question #10 and answer using your rating from <u>data element validity testing</u> – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

This was primarily done on a theoretical basis with a focus on risk adjustment.

## OVERALL RELIABILITY RATING

37. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

My overall rating reflects concerns about the ACO measure – the original readmission measure was focused on a single institution (the hospital). As the measure gets extended to ACOs, there are new issues about changing patient enrollment status over time and other issues that could theoretically affect reliability.

Also, the authors refer the reader to supporting documentation on the results of item level reliability testing – the approach is well documented, but it is difficult to find specific results for the data used for this reliability and validity testing. The measure developers state there are no concerns, but it would helpful to more easily inspect the actual results.

# VALIDITY

## Assessment of Threats to Validity

38. Were potential threats to validity that are relevant to the measure empirically assessed ()?

**REFERENCE:** Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

An all-cause risk adjusted readmission measure is a complex construct. On the one hand you have the events itself – did a readmission occur after an inpatient stay? The measure developers carefully consider planned versus unplanned admissions after hospitalization. The second component is the risk adjustment, designed basically to make the comparison between hospitals fair by adjusting for clinical variation at the facility level. This is the bulk of the validity testing (both face validity and empirical testing) work shown by the authors.

39. Analysis of potential threats to validity: Any concerns with measure exclusions?

**REFERENCE:** Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

I am always concerned about exclusions – the measure developers do show the distribution of exclusions which is very helpful. The rates look very low, suggesting this should not materially affect a hospital or ACOs score. Medical Treatment of Cancer is the largest driver of exclusions – whether or not this is appropriate seems like a clinical decision.

40. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)
**REFERENCE:** Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☒Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?
***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*
☐Yes (please explain below then go to Question #14)
☒No (go to Question #14)
☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

The measure developers describe a hybrid model development process considering earlier, condition specific models along with other candidate risk adjuster. Items are dropped from the model based on a mix of statistical and clinical characteristics. They do not include severity markers from the inpatient stay which is a plus. As the developers point out, these are potential indicators of inpatient quality. They are also 'gamble' since they are within the control of the inpatient provider. Of course, this would be said for the discharge condition category as well. However, this should be supported by documentation in the EHR. For AMI, heart failure and pneumonia the measure developers compared claims based risk adjusters to severity markers derived from chart review.

I do question the use of African-American race as a socio-demographic factor in the hospital risk model, particularly since is appears to have little impact on the c-statistic and overall model fit. The literature based argument is not compelling despite the depth of the review. SES and Dual-eligible, on the other hand, seem appropriate based on the evidence presented in the application.

41. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
**REFERENCE:** Testing attachment, section 2b4.
☒Yes (please explain below then go to Question #15)
☐No (go to Question #15)

The measure developers test construct validity by comparing hospital and ACO readmission rates for those with high performance on other outcomes measures. It is not clear conceptually why readmissions would specifically be related to these rankings. In fact, the measure developers raise this same question. In addition to this issue, I also have some concerns about un-measured severity differences between hospitals and ACOs. Although the measure developers think very carefully about the risk model, there are simply limits to what can be identified in claims data.

42. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
**REFERENCE:** Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☐No (go to Question #16)
☒Not applicable (go to Question #16)

43. Analysis of potential threats to validity: Any concerns regarding missing data?
**REFERENCE:** Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

## Assessment of Measure Testing

44. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
**REFERENCE:** Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☒Yes (go to Question #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

45. Was validity testing conducted with computed performance measure scores for each measured entity?
**REFERENCE:** Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

46. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

47. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☒Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

I am not convinced that the use of other quality and outcome measures at the hospital or ACO level provides definitive evidence of validity. Rather, it points in positive direction. My primary concern is the correlation between readmissions measures and totally unrelated constructs like screening or vaccination (this is seen in the ACO measure). The measure developers address this for the ACO measure by saying, 'Quality across even unrelated clinical area or processes of care might be weakly correlated because they all reflect the global quality of the measured ACO.' This could be true, or it could simply be spurious correlation. It is always difficult to look at the relationship among many measures and understand the true underlying relationship. Instead, the measure developers may have been better served by focusing narrowly on things like discharge planning or other things that are known to have a strong relationship with readmissions.

48. Was validity testing conducted with patient-level data elements?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

49. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

*Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

50. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

51. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

**NOTE:** If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

52. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

## OVERALL VALIDITY RATING

53. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

The reliability and validity testing of this measure was very thoughtful and through. However, readmissions is a deceptively complex phenomena. It would be helpful for the measure developers to share more of their thinking about the underlying construct and how you make fair comparisons across hospitals or ACOs. This would likely lead to some additional validity tests. That said, this seems like a model application that could be helpful for other measure developers.

# FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

54. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

**REFERENCE:** Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

# Measure Number: 1789
# Measure Title: Hospital-Wide Readmission Measure (HWR)

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

**Instructions for filling out this form:**

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

## RELIABILITY

55. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

**REFERENCE:** "MIF_xxxx" document

***NOTE****: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

***TIPS****: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

56. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

**REFERENCE:** "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3) At the hospital level

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

57. Was reliability testing conducted with computed performance measure scores for each measured entity?

**REFERENCE**: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

58. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☐Yes (go to Question #5)

☒No (please explain below, then go to question #5 and rate as INSUFFICIENT)

The developers report ICC as 0.80 in their split test reliability for the hospital level data and 0.62 for the ACO data. By conventional interpretation, this level of agreement is "substantial" at the hospital level and good at the ACO level [testing attachment page 12]

But the testing attachment also presents the cut points for penalties and rewards for ACOs and being classified better, worse, or no different from the national average [testing attachment page 43 and 44]. The ACO penalties are tightly clustered, with small changes in scores resulting in real changes in payment. The range from the 30$^{th}$ to 90$^{th}$ percentile, associated with a 2 point bonus in payment, is 0.72 percent and the 30-90$^{th}$ percentile range is 15.32-14.54.

The question is therefore not whether the ICC is by some standard in the literature "substantial" or "good" but whether the split scores are sufficiently tight that the relative percentile in which a hospital or ACO is classified is relatively stable, so that the assessment of above, at or below average, or percentile in the distribution doesn't change for most facilities, or the shift in decile doesn't change by many deciles. We can't assess this from the data presented. Before assessing this measure as reliable, I would like to see data on the stability of

the rankings and the extent to which institutions in the split sample analysis change deciles and by how much.

I did a quick simulation of a data set with an ICC of 0.8, a higher level of agreement than was observed for the ACO data, and with approximately 5% of the sample classified as above and a similar percentage classified as below, more than half the sample classified as above or below in the first sample were in the middle group in the second, with an equivalent number that were in the middle group moving to the above or below groups.

[In the same simulated data, less than half the hospitals stayed in their original decile, approximately 30% shifted one decile, 15% two deciles, and 10% three or more deciles. Ideally, we would have this data directly from the developers. While the hospital measure is not the focus of this review, I would be concerned about the reliability at this level.]

Given that penalties or bonuses are being made in payment and the simulation suggests that half the hospitals being penalized or rewarded change in the above, below or no different from average scenario and bonuses based on deciles can be shifted 2 steps for a quarter of the hospitals, I am not comfortable assessing this measure for ACOs with its ICC of 0.62 as providing sufficient reliability for its intended use.

59. **RATING (score level)** - What is the level of certainty or confidence that the <u>performance measure scores</u> are reliable?

    **REFERENCE:** Testing attachment, section 2a2.2

    ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

    ☐High (go to Question #6)

    ☐Moderate (go to Question #6)

    ☐Low (please explain below then go to Question #6)

    ☒Insufficient (go to Question #6)

60. Was reliability testing conducted with <u>patient-level data elements</u> that are used to construct the performance measure?

    **REFERENCE:** Testing attachment, section 2a2.

    ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

    ☒Yes (go to Question #7) See response to question 7 for modifications of this "Yes"

    ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

61. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

*Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☒Yes (go to Question #8) The quality of the data elements for this measure are not formally tested, which should result in a NO, and rating of insufficient. The measure is, however, based on claims data, as are many measures, and the developers discuss decisions about choosing measures that are more reliable and audited periodically by CMS, and testing comparing some chart reviewed vs. claims based measures. A decision to accept claims based data as sufficiently reliable appears to me to have been made as a matter of policy and practice long ago, and I am basically supportive of this decision.

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

62. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

**REFERENCE:** Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

63. Was **empirical VALIDITY testing** of patient-level data conducted?

**REFERENCE:** testing attachment section 2b1.

**NOTE:** Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

## OVERALL RELIABILITY RATING

64. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise,

unambiguous, and complete]

☒Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

See response to question 4. The ICC for the split sample tests, while high in terms of standards in the literature, may not be sufficiently high to assure that the rankings and quintiles of the scores are sufficiently stable from sample to sample.

# VALIDITY

## Assessment of Threats to Validity

65. Were potential threats to validity that are relevant to the measure empirically assessed ()?

**REFERENCE:** Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

66. Analysis of potential threats to validity: Any concerns with measure exclusions?

**REFERENCE:** Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

***BUT standing committee should review list of planned readmissions for appropriateness and completeness.

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

67. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

**REFERENCE:** Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social*

*risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for* ***not risk adjusting*** *provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☒Yes (please explain below then go to Question #14)

***The clinical risk adjustment model appears to be robust. C-stats not great but reasonable. Social determinants as measured (%African American, %Dual, zip code level SES variables) don't explain much variance. There is an ongoing discussion in the standing committees about whether CMS should be aggressively trying to improve these (e.g., census tract measures of SES rather than zip, based on beneficiary addresses), but there is no strong basis for arguing that risk adjustment should be made using these measures.

☐No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

68. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
**REFERENCE:** Testing attachment, section 2b4.
☐Yes (please explain below then go to Question #15)
☒No (go to Question #15)

69. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
**REFERENCE:** Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☒No (go to Question #16)
☐Not applicable (go to Question #16)

70. Analysis of potential threats to validity: Any concerns regarding missing data?
**REFERENCE:** Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

## Assessment of Measure Testing

71. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

72. Was validity testing conducted with computed performance measure scores for each measured entity?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

73. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

74. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☒Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

75. Was validity testing conducted with patient-level data elements?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

Based on prior review of data elements for other measures, or reliance on audit. Not directly assessed for this measure

76. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:*** *Testing attachment, section 2b1.*

***TIPS****: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

*Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

77. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

78. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

**NOTE:** If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

**REFERENCE:** Testing attachment, section 2b1.

***TIPS****: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

79. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS****: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

### OVERALL VALIDITY RATING

80. **OVERALL RATING OF VALIDITY** taking into account the results and scope of <u>all</u> testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

The underlying logic of this measure is that if unplanned readmissions are higher than expected, the "excess" readmissions are avoidable through better pre-discharge planning, patient education or linkage/communication with post-acute services. The evidence of reduction in readmissions cited provides some support, but the substantive committee should discuss whether there is a need for further testing of this assumption and premise of the measure.

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there <u>are</u> threats to validity and/or threats to validity were <u>not assessed</u>]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level <u>is required</u>; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

## FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

81. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

**REFERENCE:** Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

# Measure Number: 1789
# Measure Title: Hospital-Wide All-Cause Unplanned Readmission (HWR)

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

**Instructions for filling out this form:**

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

## RELIABILITY

82. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

**REFERENCE:** "MIF_xxxx" document

***NOTE****: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

***TIPS****: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

83. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

**REFERENCE:** "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS****: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

84. Was reliability testing conducted with computed performance measure scores for each measured entity?

**REFERENCE**: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS****: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

85. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS****: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

86. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS****: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

87. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

**REFERENCE:** Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

88. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

*Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☒Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

89. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

**REFERENCE:** Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

90. Was **empirical VALIDITY testing** of patient-level data conducted?

**REFERENCE:** testing attachment section 2b1.

**NOTE:** Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

## OVERALL RELIABILITY RATING

91. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

# VALIDITY

## Assessment of Threats to Validity

92. Were potential threats to validity that are relevant to the measure empirically assessed ()?

**REFERENCE:** Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

93. Analysis of potential threats to validity: Any concerns with measure exclusions?

**REFERENCE:** Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

94. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

**REFERENCE:** Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☒Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)?*

*Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

95. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

**REFERENCE:** Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

96. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

**REFERENCE:** Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☒No (go to Question #16)

☐Not applicable (go to Question #16)

97. Analysis of potential threats to validity: Any concerns regarding missing data?

**REFERENCE:** Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

## Assessment of Measure Testing

98. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

99. Was validity testing conducted with computed performance measure scores for each measured entity?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

100. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

101. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☒Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

102. Was validity testing conducted with patient-level data elements?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

103. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

*Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

104. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

105. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

**NOTE:** If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

106. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

## OVERALL VALIDITY RATING

107. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

## FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

108. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

**REFERENCE:** Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

# Measure Number: 1789
# Measure Title: Hospital-Wide All-Cause Unplanned Readmission (HWR)

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

**Instructions for filling out this form:**

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

## RELIABILITY

109. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

**REFERENCE:** "MIF_xxxx" document

***NOTE****: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

***TIPS****: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

110. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

**REFERENCE:** "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS****: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

111. Was reliability testing conducted with computed performance measure scores for each measured entity?

**REFERENCE**: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS****: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

112. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS****: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

113. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS****: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☒High (go to Question #6)

☐Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

114. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

**REFERENCE:** Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

115. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

**REFERENCE:** Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

*Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☒Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

116. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

**REFERENCE:** Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

117. Was **empirical VALIDITY testing** of patient-level data conducted?

**REFERENCE:** testing attachment section 2b1.

**NOTE:** Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

## OVERALL RELIABILITY RATING

118. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

# VALIDITY

## Assessment of Threats to Validity

119. Were potential threats to validity that are relevant to the measure empirically assessed ()?

**REFERENCE:** Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

120. Analysis of potential threats to validity: Any concerns with measure exclusions?

**REFERENCE:** Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

121. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

**REFERENCE:** Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☒Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)?*

*Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

122. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

**REFERENCE:** Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

123. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

**REFERENCE:** Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☒No (go to Question #16)

☐Not applicable (go to Question #16)

124. Analysis of potential threats to validity: Any concerns regarding missing data?

**REFERENCE:** Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

## Assessment of Measure Testing

125. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

126. Was validity testing conducted with computed performance measure scores for each measured entity?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

127. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

128. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☒High (go to Question #21)

☐Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

129. Was validity testing conducted with patient-level data elements?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

130. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

*Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)*

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

131. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

132. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

**NOTE:** If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

133. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

**REFERENCE:** Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

## OVERALL VALIDITY RATING

134. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

## FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

135. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

**REFERENCE:** Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

## NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

**Measure Number** (*if previously endorsed*)**:** 1789

**Measure Title**: Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)

**IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:** N/A

**Date of Submission**: 1/29/2016

**Instructions**

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

**Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.**

### 1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

**Notes**

**3.** Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

**4.** The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

**5.** Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

**6.** Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

**1a.1.This is a measure of**: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Health outcome: 30-day, hospital-wide, all-cause, unplanned readmission

☐Patient-reported outcome (PRO): Click here to name the PRO

*PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors*

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

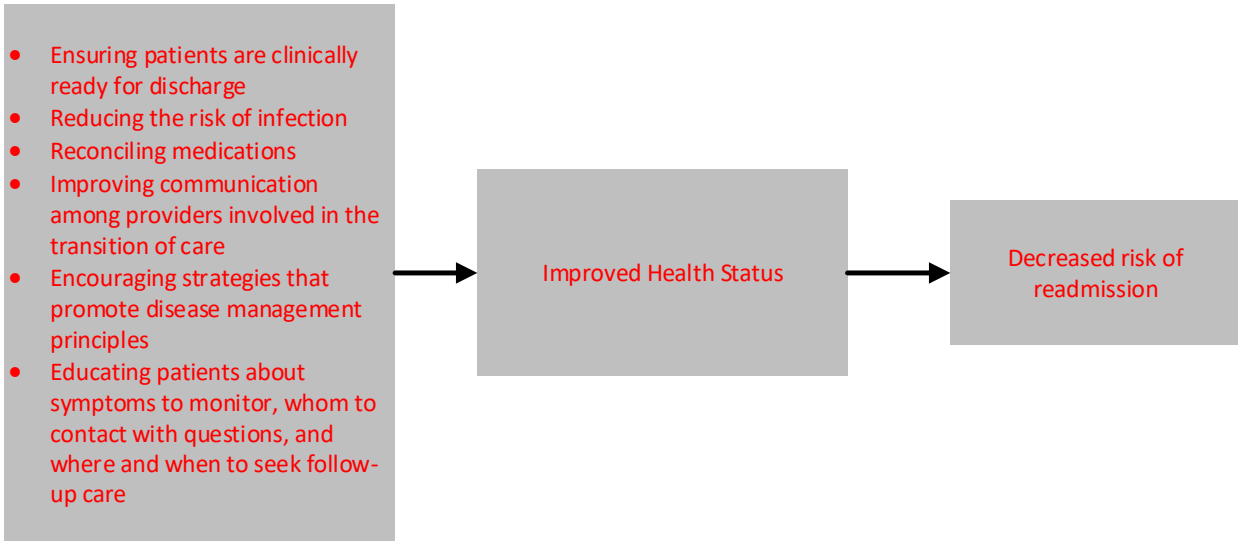
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

---

**HEALTH OUTCOME/PRO PERFORMANCE MEASURE** *If not a health outcome or PRO, skip to 1a.3*

**1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.**




The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, risk-standardized readmission rates of unplanned, all-cause readmission after admission for any eligible condition within 30 days of hospital discharge. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of, and response to, complications, patient safety and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk-adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This readmission measure was developed to identify institutions, whose performance is better or worse than would be expected based on their patient case-mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

**1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).**

The diagram above indicates some of the many care processes that can influence readmission risk. In general, randomized controlled trials have shown that improvement in the following areas can directly reduce readmission rates: quality of care during the initial admission; improvement in communication with patients, their caregivers, and their clinicians; patient education; predischarge assessment; and coordination of care after discharge. Evidence that hospitals have been able to reduce readmission rates through these quality-of-care initiatives illustrates the degree to which hospital practices can affect readmission rates. Successful randomized trials have reduced 30-day readmission rates by 20-40% [1-11]. Since 2008, 14 Medicare Quality Improvement Organizations have been funded to focus on care transitions, applying lessons learned from clinical trials. Several have been notably successful in reducing readmissions. The strongest evidence supporting the efficacy of improved discharge processes

and enhanced care at transitions is a randomized controlled trial by the Project RED (Re-Engineered Discharge) intervention, in which a nurse was assigned to each patient as a discharge advocate, responsible for patient education, follow-up, medication reconciliation, and preparing individualized discharge instructions sent to the patient's primary care provider and there was a follow-up phone call from a pharmacist within 4 days of discharge demonstrated a 30% reduction in 30-day readmissions [1]. Hospital processes that reflect the quality of inpatient and outpatient care such as discharge planning, medication reconciliation, and coordination of outpatient care have been shown to reduce readmission rates [12]. Although readmission rates are also influenced by hospital system characteristics, such as the bed capacity of the local health care system, these hospital characteristics should not influence quality of care [13]. Therefore, this measure does not risk adjust for such hospital characteristics.

Studies have estimated the rate of preventable readmissions to be as low as 12% and as high as 76% [14, 15]. Given that studies have shown readmissions to be related to quality of care, and that interventions have been able to reduce 30-day readmission rates, it is reasonable to consider an all-condition readmission rate as a quality measure.

The hospital-wide risk-standardized readmission rate (RSRR) measure is thus intended to inform quality-of-care improvement efforts, as individual process-based performance measures cannot encompass all the complex and critical aspects of care within a hospital that contribute to patient outcomes. As a result, many stakeholders, including patient organizations, are interested in outcomes measures that allow patients and providers to assess relative outcomes performance for hospitals

References:

1. Jack BW, Chetty VK, Anthony D, Greenwald JL, Sanchez GM, Johnson AE, et al. A reengineered hospital discharge program to decrease rehospitalization: a randomized trial. Ann Intern Med 2009;150(3):178-87.

2. Coleman EA, Smith JD, Frank JC, Min SJ, Parry C, Kramer AM. Preparing patients and caregivers to participate in care delivered across settings: the Care Transitions Intervention. J Am Geriatr Soc 2004;52(11):1817-25.

3. Courtney M, Edwards H, Chang A, Parker A, Finlayson K, Hamilton K. Fewer emergency readmissions and better quality of life for older adults at risk of hospital readmission: a randomized controlled trial to determine the effectiveness of a 24-week exercise and telephone follow-up program. J Am Geriatr Soc 2009;57(3):395-402.

4. Garasen H, Windspoll R, Johnsen R. Intermediate care at a community hospital as an alternative to prolonged general hospital care for elderly patients: a randomised controlled trial. BMC Public Health 2007;7:68.

5. Koehler BE, Richter KM, Youngblood L, Cohen BA, Prengler ID, Cheng D, et al. Reduction of 30-day postdischarge hospital readmission or emergency department (ED) visit rates in high-risk elderly medical patients through delivery of a targeted care bundle. J Hosp Med 2009;4(4):211-218.

6. Mistiaen P, Francke AL, Poot E. Interventions aimed at reducing problems in adult patients discharged from hospital to home: a systematic metareview. BMC Health Serv Res 2007;7:47.

7. Naylor M, Brooten D, Jones R, Lavizzo-Mourey R, Mezey M, Pauly M. Comprehensive discharge planning for the hospitalized elderly. A randomized clinical trial. Ann Intern Med 1994;120(12):999-1006.

8. Naylor MD, Brooten D, Campbell R, Jacobsen BS, Mezey MD, Pauly MV, et al. Comprehensive discharge planning and home follow-up of hospitalized elders: a randomized clinical trial. Jama 1999;281(7):613-20.

9. van Walraven C, Seth R, Austin PC, Laupacis A. Effect of discharge summary availability during post-discharge visits on hospital readmission. J Gen Intern Med 2002;17(3):186-92.

10. Weiss M, Yakusheva O, Bobay K. Nurse and patient perceptions of discharge readiness in relation to postdischarge utilization. Med Care 2010;48(5):482-6.

11. Krumholz HM, Amatruda J, Smith GL, et al. Randomized trial of an education and support intervention to prevent readmission of patients with heart failure. J Am Coll Cardiol. Jan 2 2002;39(1):83-89.

12. Nelson EA, Maruish ME, Axler JL. Effects of Discharge Planning and Compliance With Outpatient Appointments on Readmission Rates. Psychiatr Serv. July 1 2000;51(7):885-889.

13. Fisher ES, Wennberg JE, Stukel TA, Sharp SM. Hospital Readmission Rates for Cohorts of Medicare Beneficiaries in Boston and New Haven. New England Journal of Medicine. 1994;331(15):989-995.

14. Benbassat J, Taragin M. Hospital readmissions as a measure of quality of health care: advantages and limitations. Archives of Internal Medicine 2000;160(8):1074-81.

15. Medicare Payment Advisory Commission (U.S.). Report to the Congress promoting greater efficiency in Medicare. Washington, DC: Medicare Payment Advisory Commission, 2007.

*Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.*

---

**INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE**

**1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes**. Include all the steps between the measure focus and the health outcome.

N/A. This measure is not an intermediate outcome, process, or structure performance measure.

**1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?**

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

| |
|---|
| N/A |

*Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.*

---

**1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION**

**1a.4.1. Guideline citation** (*including date*) and **URL for guideline** (*if available online*):

| |
|---|
| N/A |

**1a.4.2. Identify guideline recommendation number and/or page number** and **quote verbatim, the specific guideline recommendation**.

| |
|---|
| N/A |

**1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:**

| |
|---|
| N/A |

**1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.** (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

| |
|---|
| N/A |

**1a.4.5. Citation and URL for methodology for grading recommendations** (*if different from 1a.4.1*)**:**

| |
|---|
| N/A |

**1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?**

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

---

### 1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

**1a.5.1. Recommendation citation** (*including date*) and **URL for recommendation** (*if available online*):

| N/A |
|---|

**1a.5.2. Identify recommendation number and/or page number** and **quote verbatim, the specific recommendation**.

| N/A |
|---|

**1a.5.3. Grade assigned to the quoted recommendation with definition of the grade**:

| N/A |
|---|

**1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.** (*Note: the grading system for the evidence should be reported in section 1a.7.*)

| N/A |
|---|

**1a.5.5. Citation and URL for methodology for grading recommendations** (*if different from 1a.5.1*)**:**

| N/A |
|---|

***Complete section 1a.7***

---

### 1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

**1a.6.1. Citation** (*including date*) and **URL** (*if available online*):

N/A

**1a.6.2. Citation and URL for methodology for evidence review and grading** (*if different from 1a.6.1*)**:**

N/A

***Complete section 1a.7***

---

## 1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

*If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.*

**1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?**

N/A

**1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade**: .

N/A

**1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.**

N/A

**1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range**: Click here to enter date range

N/A

**QUANTITY AND QUALITY OF BODY OF EVIDENCE**

**1a.7.5. How many and what type of study designs are included in the body of evidence**? (*e.g., 3 randomized controlled trials and 1 observational study*)

N/A

**1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence**? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

N/A

**ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE**

**1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence**? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

N/A

**1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?**

N/A

**UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE**

**1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review**.

N/A

---

## 1a.8 OTHER SOURCE OF EVIDENCE

*If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.*

**1a.8.1 What process was used to identify the evidence?**

N/A

**1a.8.2. Provide the citation and summary for each piece of evidence.**

| |
|---|
| N/A |


## Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

### Brief Measure Information

**NQF #:** 1789
**Corresponding Measures:**
De.2. **Measure Title:** Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)
Co.1.1. **Measure Steward:** Centers for Medicare & Medicaid Services (CMS)
De.3. **Brief Description of Measure:** For the hospital-wide readmission (HWR) measure that was previously endorsed and is used in the Hospital Inpatient Quality Reporting Program (IQR), the measure estimates a hospital-level risk-standardized readmission rate (RSRR) of unplanned, all-cause readmission after admission for any eligible condition within 30 days of hospital discharge. The measure reports a single summary RSRR, derived from the volume-weighted results of five different models, one for each of the following specialty cohorts based on groups of discharge condition categories or procedure categories: surgery/gynecology; general medicine; cardiorespiratory; cardiovascular; and neurology, each of which will be described in greater detail below. The measure also indicates the hospital-level standardized risk ratios (SRR) for each of these five specialty cohorts. The outcome is defined as unplanned readmission for any cause within 30 days of the discharge date for the index admission (the admission included in the measure cohort). A specified set of planned readmissions do not count in the readmission outcome. CMS annually reports the measure for patients who are 65 years or older, are enrolled in fee-for-service (FFS) Medicare, and hospitalized in non-federal hospitals.

For the All-Cause Readmission (ACR) measure version used in the Shared Savings Program (SSP), the measure estimates an Accountable Care Organization (ACO) facility-level RSRR of unplanned, all-cause readmission after admission for any eligible condition within 30 days of hospital discharge. The ACR measure is calculated using the same five specialty cohorts and estimates an ACO-level standardized risk ratio for each. CMS annually reports the measure for patients who are 65 years or older, are enrolled in FFS Medicare and are ACO assigned beneficiaries.
1b.1. **Developer Rationale:** The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, ACOs, and policy makers with information about risk-standardized all-cause unplanned readmission rates among Medicare beneficiaries 65 years and older admitted to all non-federal US acute care hospitals. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions' whose performance is better or worse than would be expected based on their patient case mix and hospital service mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

Hospital-wide readmission is a priority area for outcomes measure development as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting readmission rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices, as well as increase transparency for consumers.

For the ACR measure, several ACOs have shared with CMS the interventions they have implemented to reduce hospital readmissions. ACOs are redesigning care to improve results on the ACR measure. Some specific examples include:

1. Care coordination focusing on transitions or special populations
One ACO works to prevent readmissions to the hospital through the Transitions of Care program. A medical assistant care transition navigator conducts telephone outreach to patients at 48 hours and two weeks post-discharge.

Another ACO focuses on reducing readmissions via a home connection program for high-risk populations. Key components include ensuring a physician follow-up appointment is scheduled before discharge, ensuring patients have a personal contact for urgent needs, and ensuring patients understand how to manage their medications. Many ACOs focus on improved transitions of care for patients with end-stage renal disease to prevent readmissions.

2. Pharmacy involvement: Strategies include medication reconciliation as well as data integration with labs and pharmacies. Another strategy is increased pharmacist involvement in transitions of care: One ACO has a pharmacist focusing on transitions of care to reduce readmissions for patients with heart failure, Chronic Obstructive Pulmonary Disease (COPD), and pneumonia.

S.4. **Numerator Statement:** The outcome for the HWR measure is 30-day readmission. We define readmission as an inpatient admission for any cause, with the exception of certain planned readmissions, within 30 days from the date of discharge from an eligible index admission. If a patient has more than one unplanned admission (for any reason) within 30 days after discharge from the index admission, only one is counted as a readmission. The measure looks for a dichotomous yes or no outcome of whether each admitted patient has an unplanned readmission within 30 days. However, if the first readmission after discharge is considered planned, any subsequent unplanned readmission is not counted as an outcome for that index admission because the unplanned readmission could be related to care provided during the intervening planned readmission rather than during the index admission.

The outcome for the ACR measure is also 30-day readmission. The outcome is defined identically to what is described above for the HWR measure.

S.7. **Denominator Statement:** The measure at the hospital level includes admissions for Medicare beneficiaries who are 65 years and older and are discharged from all non-federal, acute care inpatient US hospitals (including territories) with a complete claims history for the 12 months prior to admission.

The measure at the ACO level includes all relevant admissions for ACO assigned beneficiaries who are 65 and older and are discharged from all non-Federal short-stay acute care hospitals, including critical access hospitals.

Additional details are provided in S.9 Denominator Details.

S.10. **Denominator Exclusions:** The measure excludes index admissions for patients:

1. Admitted to Prospective Payment System (PPS)-exempt cancer hospitals;
2. Without at least 30 days post-discharge enrollment in FFS Medicare;

3. Discharged against medical advice (AMA);
4. Admitted for primary psychiatric diagnoses;
5. Admitted for rehabilitation; or
6. Admitted for medical treatment of cancer.

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Claims (Only)
S.26. **Level of Analysis:** Facility, Integrated Delivery System

**IF Endorsement Maintenance – Original Endorsement Date:** Apr 24, 2012 **Most Recent Endorsement Date:** Dec 09, 2016

**IF this measure is included in a composite, NQF Composite#/title:**

**IF this measure is paired/grouped, NQF#/title:**

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

## 1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

**1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form**
NQF_1789_HWR_NQF_Evidence_Attachment_02-15-16_v1.0.docx

**1b. Performance Gap**

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, ACOs, and policy makers with information about risk-standardized all-cause unplanned readmission rates among Medicare beneficiaries 65 years and older admitted to all non-federal US acute care hospitals. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions' whose performance is better or worse than would be expected based on their patient case mix and hospital service mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

Hospital-wide readmission is a priority area for outcomes measure development as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting readmission rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The

measure will also provide patients with information that could guide their choices, as well as increase transparency for consumers.

For the ACR measure, several ACOs have shared with CMS the interventions they have implemented to reduce hospital readmissions. ACOs are redesigning care to improve results on the ACR measure. Some specific examples include:

1. Care coordination focusing on transitions or special populations
One ACO works to prevent readmissions to the hospital through the Transitions of Care program. A medical assistant care transition navigator conducts telephone outreach to patients at 48 hours and two weeks post-discharge.

Another ACO focuses on reducing readmissions via a home connection program for high-risk populations. Key components include ensuring a physician follow-up appointment is scheduled before discharge, ensuring patients have a personal contact for urgent needs, and ensuring patients understand how to manage their medications. Many ACOs focus on improved transitions of care for patients with end-stage renal disease to prevent readmissions.

2. Pharmacy involvement: Strategies include medication reconciliation as well as data integration with labs and pharmacies. Another strategy is increased pharmacist involvement in transitions of care: One ACO has a pharmacist focusing on transitions of care to reduce readmissions for patients with heart failure, Chronic Obstructive Pulmonary Disease (COPD), and pneumonia.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The following results display the RSRRs for the HWR measure in calendar years 2011, 2012, 2013 and 2014.

Distribution of RSRRs for the HWR measure over Different Time Periods
Results for each data year
Characteristic//07/2011-06/2012//07/2012-06/2013//07/2013-06/2014/
Number of Hospitals/4,821/ /4,794/ /4,772/
Number of Admissions/7,678,216/ /7,279,853/ /6,843,808/
Mean (SD)/16.2(1.1)/15.6(0.92)/ /15.5 (0.8)/
Range (min. – max.)/10.9-22.6/ /11.0-21.4/ /11.4-20.1/
Minimum/10.9/ /11.0/ /11.4/
10th percentile/15.1/ / 14.6/ /14.6/
20th percentile/15.4/ /14.9/ /14.9/
30th percentile/15.7/ /15.2/ /15.1/
40th percentile/15.9/ /15.4/ /15.3/
50th percentile/16.1/ /15.5/ /15.4/
60th percentile/16.4/ /15.7/ /15.6/
70th percentile/16.6/ /15.9/ /15.8/
80th percentile/17.0/ /16.2/ /16.0/
90th percentile/17.5/ /16.8/ /16.5/
Maximum/22.6/ /21.4/ /20.1/

The following results display the RSRRs for the ACR measure in calendar years 2013, 2014, and 2015.

Distribution of RSRRs for the ACR measure over Different Time Periods

Results for each data year
Characteristic//01/2013-12/2013//01/2014-12/2014//01/2015-12/2015/
Number of ACOs/243//360//416/
Number of Admissions/915,855//1,311,746//1,721,598/
Mean (SD)/14.9(.72)//15.2(.76)//14.9(.70)/
Range (min.-max.)/13.3-18.0//13.2-18.1//13.1-17.5/
Minimum/13.3//13.2//13.1/
10th percentile/14.0//14.3//14.0/
20th percentile/14.3//14.6//14.3/
30th percentile/14.5//14.8//14.5/
40th percentile/14.7//14.9//14.6/
50th percentile/14.8//15.1//14.8/
60th percentile/15.0//15.3//15.0/
70th percentile/15.2//15.4//15.2/
80th percentile/15.4//15.7//15.5/
90th percentile/15.8//16.2//15.7/
Maximum/18.0//18.1//17.5/

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
HWR

We do not expect this measure to be "topped out" for hospitals. The following results display the RSRRs for the HWR measure.

Distribution of RSRRs for the HWR measure by Proportion of Dual-Eligible Patients:
Dates of Data: July 2013 through June 2014
Data Source: Medicare FFS claims

Characteristic//Hospitals with a low proportion (=9.8%) Dual Eligible patients//Hospitals with a high proportion (=22.6%) Dual Eligible patients
Number of Measured Hospitals// 1,257 // 1,219
Number of Patients// 2,137,895 patients in low-proportion hospitals // 927,007 in high-proportion hospitals
Maximum// 18.7 // 20.1
90th percentile// 16.2 // 16.8
75th percentile// 15.7 // 16.0
Median (50th percentile)// 15.3 // 15.6
25th percentile// 14.8 // 15.2
10th percentile// 14.3 // 14.9
Minimum // 11.5 // 12.2

Distribution of RSRRs for the HWR measure by Proportion of African-American Patients:
Dates of Data: July 2013 through June 2014
Data Source: Medicare FFS claims

Characteristic// Hospitals with a low proportion (=2.2%) African-American patients//Hospitals with a high proportion (=9.4%) African-American patients
Number of Measured Hospitals// 1,156 // 1,180
Number of Patients// 222,648 patients in low-proportion hospitals/ 2,294,715 in high-proportion hospitals
Maximum// 19.1 // 19.9
90th percentile// 16.0 // 17.1
75th percentile// 15.6 // 16.3
Median (50th percentile)// 15.4 // 15.7
25th percentile// 15.1 // 15.2
10th percentile// 14.8 // 14.8
Minimum // 12.9 // 12.2

Distribution of RSRRs for the HWR measure by Proportion of Patients with AHRQ SES Index Scores Below 45.0:
Dates of Data: July 2013 through June 2014
Data Source: Medicare FFS claims and the American Community Survey (2008-2012) data

Characteristic//Hospitals with a low proportion of patients below AHRQ SES index score of 45.0 (=5.0%)// Hospitals with a high proportion of patients below AHRQ SES index score of 45.0 (=57.1%)
Number of Measures Hospitals// 1,209 // 1,217
Number of Patients// 1,651,852 patients in hospitals with low proportion of patients below AHRQ SES index score of 45.0 //795,899 patients in hospitals with high proportion of patients below AHRQ SES index score of 45.0
Maximum// 19.9 // 20.1
90th percentile// 16.2 // 16.6
75th percentile// 15.7 // 16.0
Median (50th percentile)// 15.3 // 15.5
25th percentile// 14.9 // 15.2
10th percentile// 14.5 // 14.8
Minimum // 11.5 // 13.0

ACR

We do not expect this measure to be “topped out” for ACOs. The following results display the RSRRs for the ACR measure.

Distribution of RSRRs for the ACR measure by Proportion of Dual-Eligible Patients
Dates of Data: January 1, 2015- December 31, 2015
Data Source: Medicare FFS claims for ACO assigned/aligned beneficiaries.
Characteristic//ACOs with a low proportion (=5.1%) Dual Eligible patients// ACOs with a high proportion (=13.3%) Dual Eligible patients
Number of ACOs// 103// 104
Number of Patients// 247,252 in low-proportion of ACOs // 217,145 in high-proportion ACOs
Maximum// 16.2 // 17.5
90th percentile// 15.4 // 16.2
75th percentile// 14.9 // 15.7
Median (50th percentile)// 14.6 // 15.3
25th percentile// 14.2 // 14.7
10th percentile// 13.9 // 14.5
Minimum// 13.1 // 13.8

Characteristic// ACOs with a low proportion of patients below median AHRQ SES index score (=31.8%)// ACOs with a high proportion of patients below median AHRQ SES index score (=66.4%)
Number of ACOs// 104 // 104
Number of patients// 336,504 // 185,644
Maximum// 17.2 // 16.7
90th percentile// 15.9 // 15.9
75th percentile// 15.4 // 15.4
Median// 14.8 // 14.9
25th percentile// 14.4 // 14.6
10th percentile// 14.1 // 14.2
Minimum // 13.1 // 13.5

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

**1c. High Priority** (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
During 2003 and 2004, almost one fifth of Medicare beneficiaries – over 2.3 million patients – were rehospitalized within 30 days of discharge from an acute care hospital (Jencks et al., 2009). Jencks et. al. estimated that readmissions within 30 days of discharge cost Medicare more than $17 billion annually (Jencks et al., 2009). A 2006 Commonwealth Fund report further estimated that if national readmission rates were lowered to the levels achieved by the top performing regions, Medicare would save $1.9 billion annually (The Commonwealth Fund, 2006). In a 2007 report to the Congress, the Medicare Payment Advisory Commission (MedPAC) estimated that in 2005, 17.6% of hospital patients were readmitted within 30 days of discharge and that 76% of these readmissions were potentially preventable; the average payment for a "potentially preventable" readmission was estimated at approximately $7,200 (MedPAC, 2007).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. New England Journal of Medicine 2009;360(14):1418-28.

Why Not the Best? Results from a National Scorecard on U.S. Health System Performance. Fund Report. Harrisburg, PA: The Commonwealth Fund, 2006.

Medicare Payment Advisory Commission (U.S.). Report to the Congress promoting greater efficiency in Medicare. Washington, DC: Medicare Payment Advisory Commission, 2007.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A. This measure is not a PRO-PM.

## 2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

**2a.1. Specifications** The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cardiovascular, Cardiovascular : Arrythmia, Cardiovascular : Congestive Heart Failure, Cardiovascular : Coronary Artery Disease, Cardiovascular : Coronary Artery Disease (AMI), Cardiovascular : Coronary Artery Disease (PCI), Cardiovascular : Hyperlipidemia, Cardiovascular : Hypertension, Critical Care, Endocrine, Endocrine : Diabetes, Endocrine : Thyroid Disorders, Gastrointestinal (GI), Gastrointestinal (GI) : Gall Bladder Disease, Gastrointestinal (GI) : Gastroenteritis, Gastrointestinal (GI) : Gastro-Esophageal Reflux Disease (GERD), Gastrointestinal (GI) : Peptic Ulcer, Genitourinary (GU), Genitourinary (GU) : Incontinence/pelivic floor disorders, Infectious Diseases (ID), Infectious Diseases (ID) : HIV/AIDS, Infectious Diseases (ID) : Pneumonia and respiratory infections, Infectious Diseases (ID) : Sexually Transmitted, Infectious Diseases (ID) : Tuberculosis, Liver : Viral Hepatitis, Musculoskeletal, Musculoskeletal : Falls and Traumatic Injury, Musculoskeletal : Joint Surgery, Musculoskeletal : Low Back Pain, Musculoskeletal : Osteoarthritis, Musculoskeletal : Osteoporosis, Musculoskeletal : Rheumatoid Arthritis, Neurology, Neurology : Brain Injury, Neurology : Stroke/Transient Ischemic Attack (TIA), Renal, Renal : Chronic Kidney Disease (CKD), Renal : End Stage Renal Disease (ESRD), Respiratory, Respiratory : Asthma, Respiratory : Chronic Obstructive Pulmonary Disease (COPD), Respiratory : Dyspnea, Respiratory : Pneumonia, Respiratory : Sleep Apnea, Surgery, Surgery : Cardiac Surgery, Surgery : General Surgery, Surgery : Perioperative and Anesthesia, Surgery : Thoracic Surgery, Surgery : Vascular Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
N/A

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_1789_NQF_Data_Dictionary_05-26-17_v1.0.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

Annual Updates

1. Each year we update to the most current version of the Agency for Healthcare Research and Quality Clinical Classifications Software (AHRQ CCS) software by identifying any changes from the previous version that might impact the measure.
2. In addition, we have updated the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) Hierarchical Condition Categories (HCC) map annually to capture any changes that might impact the measure's risk model. The version of the HCC map used for this measure has not been updated since 2013.

Updates by year

2017

1. Updated to add the results of testing and performance of the ACR measure.
Rationale: The ACO risk standardized ACR measure is adapted from the HWR measure. It estimates the risk-adjusted percentage of ACO assigned beneficiaries who were hospitalized and readmitted to a hospital within 30 days of discharge from the index hospital admission. Testing of the measure was completed to assess performance gap, measure score reliability and validity, and impact of SES variables on risk-adjustment.

2015

1. Respecified the measure by updated to CMS Planned Readmission Algorithm (Version 4.0).
Rationale: Version 4.0 incoprporates additional improvements made following a validation study of the algorithm using data from a medical record review. These changes required additional input from clinical experts and were, therefore, not included in the changes made in version 3.0. The changes improve the accuracy of the algorithm by more correctly classifying planned and unplanned readmissions.

2014

1. Updated to CMS Planned Readmission Algorithm (Version 3.0).
Rationale: Version 3.0 incorporates improvements made following a validation study of the algorithm using data from a medical record review. These changes improve the accuracy of the algorithm by decreasing the number of readmissions that the algorithm mistakenly designated as planned by removing two procedure categories and adding several acute diagnoses.

2013

1. Updated to CMS Planned Readmission Algorithm (Version 2.1).
Rationale: Version 2.1 incorporated improvements to the original algorithm made following an extensive review by clinical experts and stakeholder feedback submitted during the HWR measure's public comment period and 2012 dry run.

3. Removed procedure CCS 61 (Other or procedures on vessels other than head and neck) from the list of procedures qualifying an admission for the surgery cohort.
Rationale: This procedure CCS was removed from the surgical cohort because patients undergoing this procedure are typically admitted primarily for cardiovascular or medical care.

4. Modified the planned readmission algorithm handling of admissions to psychiatric and rehabilitation hospitals.
Rationale: Psych and rehab hospitals in Maryland have the same provider ID number as acute care hospitals. Therefore, readmissions are not counted if the patient has a principal diagnosis code beginning with a "V57" (indication of admission to a rehab unit) or if all three of the following criteria are met: (1) the admission being evaluated as a potential readmission has a psychiatric principal discharge diagnosis code (ICD-9 codes 290-319); (2) the index admission has a discharge disposition code to a psychiatric hospital or psychiatric unit from the index admission; and (3) the admission being evaluated as a potential readmission occurred during the same day as or the day following the index discharge. The criteria for identifying such admissions are available in the

2010 Measures Maintenance Technical Report: Acute Myocardial Infarction, Heart Failure, and Pneumonia 30-Day Risk-Standardized Readmission Measures.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
*IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.*
The outcome for the HWR measure is 30-day readmission. We define readmission as an inpatient admission for any cause, with the exception of certain planned readmissions, within 30 days from the date of discharge from an eligible index admission. If a patient has more than one unplanned admission (for any reason) within 30 days after discharge from the index admission, only one is counted as a readmission. The measure looks for a dichotomous yes or no outcome of whether each admitted patient has an unplanned readmission within 30 days. However, if the first readmission after discharge is considered planned, any subsequent unplanned readmission is not counted as an outcome for that index admission because the unplanned readmission could be related to care provided during the intervening planned readmission rather than during the index admission.

The outcome for the ACR measure is also 30-day readmission. The outcome is defined identically to what is described above for the HWR measure.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Numerator Time Window: We define the time period for readmission as within 30 days from the date of discharge of the index admission.

Denominator Time Window: This measure was developed with 12 months of data and is currently publicly reported with one year of data.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
*IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.*
The measure counts readmissions to any acute care hospital for any cause within 30 days of the date of discharge of the index admission, excluding planned readmissions as defined below.

Planned Readmission Algorithm (Version 4.0)
The Planned Readmission Algorithm is a set of criteria for classifying readmissions as planned among the general Medicare population using Medicare administrative claims data. The algorithm identifies admissions that are typically planned and may occur within 30 days of discharge from the hospital.

The Planned Readmission Algorithm has three fundamental principles:

1. A few specific, limited types of care are always considered planned (obstetric delivery, transplant surgery, maintenance chemotherapy/immunotherapy, rehabilitation);
2. Otherwise, a planned readmission is defined as a non-acute readmission for a scheduled procedure; and
3. Admissions for acute illness or for complications of care are never planned.

The algorithm was developed in 2011 as part of the Hospital-Wide Readmission measure. In 2013, CMS applied the algorithm to its other readmission measures.

The Planned Readmission Algorithm and associated code tables are attached in data field S.2b (Data Dictionary or Code Table).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The measure at the hospital level includes admissions for Medicare beneficiaries who are 65 years and older and are discharged from all non-federal, acute care inpatient US hospitals (including territories) with a complete claims history for the 12 months prior to admission.

The measure at the ACO level includes all relevant admissions for ACO assigned beneficiaries who are 65 and older and are discharged from all non-Federal short-stay acute care hospitals, including critical access hospitals.

Additional details are provided in S.9 Denominator Details.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Elderly

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
To be included in the hospital level measure, cohort patients must be:

1. Enrolled in Medicare fee-for-service (FFS) Part A for the 12 months prior to the date of admission and during the index admission;
2. Aged 65 or over;
3. Discharged alive from a non-federal short-term acute care hospital; and
4. Not transferred to another acute care facility.

The ACO version of this measure has the additional criterion that only hospitalizations for ACO-assigned beneficiaries that meet all of the other criteria listed above are included. The cohort definition is otherwise identical to that of the HWR described below.

The measure aggregates the ICD-9 principal diagnosis and all procedure codes of the index admission into clinically coherent groups of conditions and procedures (condition categories or procedure categories) using the AHRQ CCS. There are a total of 285 mutually exclusive AHRQ condition categories, most of which are single, homogenous diseases such as pneumonia or acute myocardial infarction. Some are aggregates of conditions, such as “other bacterial infections.” There are a total of 231 mutually exclusive procedure categories. Using the AHRQ CCS procedure and condition categories, the measure assigns each index hospitalization to one of five mutually exclusive specialty cohorts: surgery/gynecology, cardiorespiratory, cardiovascular, neurology, and medicine. The rationale behind this organization is that conditions typically cared for by the same team of clinicians are expected to experience similar added (or reduced) levels of readmission risk.

The measure first assigns admissions with qualifying AHRQ procedure categories to the Surgery/Gynecology Cohort. This cohort includes admissions likely cared for by surgical or gynecological teams.

The measure then sorts admissions into one of the four remaining specialty cohorts based on the AHRQ diagnosis category of the principal discharge diagnosis:

The Cardiorespiratory Cohort includes several condition categories with very high readmission rates such as pneumonia, chronic obstructive pulmonary disease, and heart failure. These admissions are combined into a

single cohort because they are often clinically indistinguishable and patients are often simultaneously treated for several of these diagnoses.

The Cardiovascular Cohort includes condition categories such as acute myocardial infarction that in large hospitals might be cared for by a separate cardiac or cardiovascular team.

The Neurology Cohort includes neurologic condition categories such as stroke that in large hospitals might be cared for by a separate neurology team.

The Medicine Cohort includes all non-surgical patients who were not assigned to any of the other cohorts.

The full list of the specific diagnosis and procedure AHRQ CCS categories used to define the specialty cohorts are attached in data field S.2b (Data Dictionary or Code Table).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
The measure excludes index admissions for patients:

1. Admitted to Prospective Payment System (PPS)-exempt cancer hospitals;
2. Without at least 30 days post-discharge enrollment in FFS Medicare;
3. Discharged against medical advice (AMA);
4. Admitted for primary psychiatric diagnoses;
5. Admitted for rehabilitation; or
6. Admitted for medical treatment of cancer.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
1. Admitted to a PPS-exempt cancer hospital, identified by the Medicare provider ID.
2. Admissions without at least 30 days post-discharge enrollment in FFS Medicare are determined using data captured in the Medicare Enrollment Database (EDB).
3. Discharges against medical advice (AMA) are identified using the discharge disposition indicator in claims data.
4. Admitted for primary psychiatric disease, identified by a principal diagnosis in one of the specific AHRQ CCS categories listed in the attached data dictionary.
5. Admitted for rehabilitation care, identified by the specific ICD-9 diagnosis codes included in CCS 254 (Rehabilitation care; fitting of prostheses; and adjustment of devices).
6. Admitted for medical treatment of cancer, identified by the specific AHRQ CCS categories listed in the attached data dictionary.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

**S.14. Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

Our approach to risk adjustment is tailored to and appropriate for a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz et al., 2006).

The HWR measure employs a hierarchical logistic regression model to create a hospital-level 30-day RSRR. In brief, the approach simultaneously models data at the patient and hospital levels to account for the variance in patient outcomes within and between hospitals (Normand & Shahian, 2007). At the patient level, the model adjusts the log-odds of readmission within 30 days of discharge for age and selected clinical covariates. At the hospital level, the approach models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of readmission at the hospital, after accounting for patient risk. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

We use a fixed, common set of variables in all our models for simplicity and ease of data collection and analysis. However, we estimate a hierarchical logistic regression model for each specialty cohort separately, and the coefficients associated with each variable may vary across specialty cohorts.

Candidate and Final Risk-adjustment Variables: Candidate variables were patient-level risk-adjustors that were expected to be predictive of readmission, based on empirical analysis, prior literature, and clinical judgment, including age and indicators of comorbidity and disease severity. For each patient, covariates are obtained from claims records extending 12 months prior to and including the index admission. For the measure currently implemented by CMS, these risk-adjusters are identified using inpatient Medicare FFS claims data.

The model adjusts for case-mix differences based on the clinical status of patients at the time of admission. We use condition categories (CCs), which are clinically meaningful groupings of more than 15,000 ICD-9-CM diagnosis codes (Pope et al., 2000). A file that contains a list of the ICD-9-CM codes and their groupings into CCs is attached in data field S.2b (Data Dictionary or Code Table). In addition, only comorbidities that convey information about the patient at admission or in the 12 months prior, and not complications that arise during the course of the index hospitalization, are included in the risk adjustment. Hence, we do not risk adjust for CCs that may represent adverse events of care when they are only recorded in the index admission. The models also include a condition-specific indicator for all AHRQ CCS categories with sufficient volume (defined as those with more than 1,000 admissions nationally each year for Medicare FFS data) as well as a single indicator for conditions with insufficient volume in each model.

The final set of risk adjustment variables are listed in the attached Data Dictionary.

Demographics
Age-65 (years, continuous) for patients aged 65 or over cohorts; or Age (years, continuous) for patients aged 18 and over cohorts

Comorbidities
Metastatic cancer or acute leukemia (CC 7)
Severe cancer (CC 8-9)
Other cancers (CC 10-12)
Severe hematological disorders (CC 44)
Coagulation defects and other specified hematological disorders (CC 46)
Iron deficiency or other unspecified anemias and blood disease (CC 47)
End-stage liver disease (CC 25-26)
Pancreatic disease (CC 32)

Dialysis status (CC 130)
Renal failure (CC 131)
Transplants (CC 128, 174)
Severe infection (CC 1, 3-5)
Other infectious diseases and pneumonias (CC 6, 111-113)
Septicemia/shock (CC 2)
Congestive heart failure (CC 80)
Coronary atherosclerosis or angina, cerebrovascular disease (CC 81-84, 89, 98-99, 103-106)
Specified arrhythmias and other heart rhythm disorders (CC 92-93)
Cardio-respiratory failure or shock (CC 79)
Chronic obstructive pulmonary disease (COPD) (CC 108)
Fibrosis of lung or other chronic lung disorders (CC 109)
Protein-calorie malnutrition (CC 21)
Disorders of fluid/electrolyte/acid-base (CC 22-23)
Rheumatoid arthritis and inflammatory connective tissue disease (CC 38)
Diabetes mellitus (DM) or DM complications (CC 15-20, 119-120)
Decubitus ulcer or chronic skin ulcer (CC 148-149)
Hemiplegia, paraplegia, paralysis, functional disability (CC 67-69, 100-102, 177-178)
Seizure disorders and convulsions (CC 74)
Respirator dependence/tracheostomy status (CC 77)
Drug/alcohol psychosis or dependence (CC 51-52)
Psychiatric comorbidity (CC 54-56, 58, 60)
Hip fracture/dislocation (CC 158)

Principal Diagnoses
Refer to the 2015 Measure Updates and Specifications: Hospital-Wide All-Cause Unplanned Readmission - Version 4.0 referenced here for the full lists of principal diagnosis AHRQ CCS categories included in each specialty cohort risk adjustment model.

The ACR measure employs the same risk adjustment methodology and uses the same risk variables.

References:
Krumholz HM, Brindis RG, Brush JE, et al. 2006. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation 113: 456-462.

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.

Pope GC, et al. 2000. Principal Inpatient Diagnostic Cost Group Models for Medicare Risk Adjustment. Health Care Financing Review 21(3): 93-118.

**S.15. Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
*Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.*
Available in attached Excel or csv file at S.2b

**S.15a. Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
This measure estimates a hospital-level 30-day all-cause RSRR using hierarchical logistic regression models. In brief, the approach simultaneously models data at the patient, and hospital levels to account for variance in patient outcomes within and between hospitals (Normand et al., 2007). At the patient level, it models the log-odds of readmission within 30 days of discharge using age, selected clinical covariates, and a hospital -specific effect. At the hospital level, the approach models the hospital- specific effects as arising from a normal distribution. The hospital effect represents the underlying risk of a readmission, after accounting for patient risk. The hospital-specific effects are given a distribution to account for the clustering (non-independence) of patients within the same hospital (Normand et al., 2007). If there were no differences among hospitals, then after adjusting for patient risk, the hospital effects should be identical across all hospitals.

Admissions are assigned to one of five mutually exclusive specialty cohort groups consisting of related conditions or procedures. For each specialty cohort group, the standardized readmission ratio (SRR) is calculated as the ratio of the number of "predicted" readmissions to the number of "expected" readmissions at a given hospital. For each hospital, the numerator of the ratio is the number of readmissions within 30 days predicted based on the hospital's performance with its observed case mix and service mix, and the denominator is the number of readmissions expected based on the nation's performance with that hospital's case mix and service mix. This approach is analogous to a ratio of "observed" to "expected" used in other types of statistical analyses. It conceptually allows a particular hospital's performance, given its case mix and service mix, to be compared to an average hospital's performance with the same case mix and service mix. Thus, a lower ratio indicates lower-than-expected readmission rates or better quality, while a higher ratio indicates higher-than-expected readmission rates or worse quality.

For each specialty cohort, the "predicted" number of readmissions (the numerator) is calculated by using the coefficients estimated by regressing the risk factors (found in Table D.9) and the hospital-specific effect on the risk of readmission. The estimated hospital-specific effect for each cohort is added to the sum of the estimated regression coefficients multiplied by patient characteristics. The results are log transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" number of readmissions (the denominator) is obtained in the same manner, but a common effect using all hospitals in our sample is added in place of the hospital-specific effect. The results are log transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimate the model coefficients using the data in that period.

The specialty cohort SRRs are then pooled for each hospital using a volume-weighted geometric mean to create a hospital-wide composite SRR. The composite SRR is multiplied by the national observed readmission rate to produce the RSRR. The statistical modeling approach is described fully in Appendix A and in the original methodology report (Horwitz et al., 2012).

The ACR quality measure was adapted from the HWR quality measure. The unit of analysis was changed from the hospital to the ACO. This was possible because both the HWR and ACR measures assess readmission performance for a population that clusters patients together (either in hospitals or in ACOs). The goal is to isolate the effects of beneficiary characteristics on the probability that a patient will be readmitted from the

effects of being in a specific hospital or ACO. In addition, planned readmissions are excluded for the ACR quality measure in the same way that they are excluded for the HWR measure. The ACR measure is calculated identically to what is described above for the HWR measure.

References:

Horwitz L, Partovian C, Lin Z, et al. Hospital-Wide All-Cause Unplanned Readmission Measure: Final Technical Report. 2012; http://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228889825199&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DDryRun_HWR_TechReport_081012.pdf&blobcol=urldata&blobtable=MungoBlobs. Accessed 30 April, 2014.

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22(2): 206-226.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A. This measure is not based on a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A. This measure is not based on a survey or patient-reported data.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing values are rare among variables used from claims data in this measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
*If other, please describe in S.24.*
Claims (Only)

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Data sources for the Medicare FFS measure:

HWR

1. Medicare Part A claims data for calendar years 2007 and 2008 were combined and then randomly split into two equal subsets (development sample and validation sample). Risk variable selection was done using the development sample, the risk models for each of the five specialty cohorts in the measure were applied to the validation sample and the models' performance was compared. In addition we re-tested the models in Medicare Part A claims data from calendar year 2009 to look for temporal stability in the models' performance. The number of measured entities and index admissions are listed below by specialty cohort.

2. Medicare Enrollment Database (EDB): This database contains Medicare beneficiary demographic, benefit/coverage, and vital status information. This data source was used to obtain information on several inclusion/exclusion indicators such as Medicare status on admission and following discharge from index admission

ACR

1. Medicare Part A claims data for calendar years 2013, 2014, and 2015.

2. Medicare Enrollment Database (EDB).

Reference:
Fleming C., Fisher ES, Chang CH, Bubolz D, Malenda J. Studying outcomes and hospital utilization in the elderly: The advantages of a merged data base for Medicare and Veterans Affairs Hospitals. Medical Care. 1992; 30(5): 377-91.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Integrated Delivery System

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Clinician Office/Clinic, Hospital, Hospital : Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A.

**2a. Reliability – See attached Measure Testing Submission Form**
**2b. Validity – See attached Measure Testing Submission Form**
NQF_1789_ACR_NQF_Testing_Attachment_Final.Submission.20170523.docx

## NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

**Measure Number** (*if previously endorsed*)**:** 1789
**Measure Title**: Hospital-Wide Readmission Measure (HWR)
**Date of Submission**: 4/25/2017
**Type of Measure:**

| | |
|---|---|
| ☒ Outcome (*including PRO-PM*) | ☐ Composite – ***STOP – use composite testing form*** |
| ☐Intermediate Clinical Outcome | ☐ Cost/resource |
| ☐ Process | ☐ Efficiency |
| ☐ Structure | |

**Instructions**

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

**Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.**

**2a2. Reliability testing** [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

**2b2. Validity testing** [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

**2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

**AND**

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

**2b4. For outcome measures and other measures when indicated** (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

**OR**

- rationale/data support no risk adjustment/ stratification.

**2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

**OR**

there is evidence of overall less-than-optimal performance.

**2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results**.

**2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

**Notes**

**10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-

item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

**11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). 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.

**12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
**13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

**14.** Risk factors that influence outcomes should not be specified as exclusions
**15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

## 1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

| **Measure Specified to Use Data From:**<br>**(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
|---|---|
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| ☒ administrative claims | ☒ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| ☐ abstracted from electronic health record | ☐ abstracted from electronic health record |
| ☐ eMeasure (HQMF) implemented in EHRs | ☐ eMeasure (HQMF) implemented in EHRs |
| ☐ other: Click here to describe | ☒ other: Census Data/American Community Survey |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

| |
|---|
| The datasets used for testing included Medicare Part A inpatient claims and the Medicare Enrollment Database (EDB). Census as well as claims data were used to assess socioeconomic factors and race (dual-eligible and African American race variables were obtained through enrollment data; Agency for Healthcare Research and Quality (AHRQ) socioeconomic status (SES) index score obtained through census data). The dataset used varies by testing type; see Section 1.7 for details. |

**1.3. What are the dates of the data used in testing**? Click here to enter date range

| |
|---|
| The dates used vary by testing type; see Section 1.7 for details. |

**1.4. What levels of analysis were tested**? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

| **Measure Specified to Measure Performance of:**<br>**(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
|---|---|
| ☐ individual clinician | ☐ individual clinician |
| ☐ group/practice | ☐ group/practice |
| ☒ hospital/facility/agency | ☒ hospital/facility/agency |
| ☐ health plan | ☐ health plan |
| ☒ other: Accountable Care Organization | ☒ other: Accountable Care Organization |

**1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

| |
|---|
| For the HWR measure, hospitals are the measured entities. All non-federal, acute care inpatient US hospitals (including territories) with Medicare fee-for-service (FFS) beneficiaries aged 65 years and older are included. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.<br><br>For the ACR measure, ACO's are the measured entities. All non-federal, acute care inpatient US hospitals (including territories) with Medicare fee-for-service (FFS) assigned ACO beneficiaries aged 65 years and older are included. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details. |

**1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

| |
|---|
| The number of admissions/patients varies by testing type: see Section 1.7 for details |

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

**<u>Data for the HWR measure</u>**

The datasets, dates, number of measured entities and number of admissions used in each type of testing are as follows:

<u>For reliability testing (Section 2a2)</u>
The reliability of the model was tested by randomly selecting 50% of the Medicare patients aged 65 years and over within each hospital in the most recent 1-year measure cohort and calculating the measure results for each hospital. We then calculated the measure results for the remaining 50% of patients within each hospital and compare the two. Thus, for reliability testing, we randomly split **Dataset 1** into two samples. In each year of measure reevaluation, we re-fit the model and examine frequencies and model coefficients of risk variables (condition categories for patient comorbidities) and model fit in the new year of data (**Dataset 1** below).

**Dataset 1** (2015 public reporting cohort version 4.0): Medicare Part A Inpatient Claims and Medicare Enrollment Database
Dates of Data: July 1, 2013 – June 30, 2014
Number of index admissions: 6,843,808

Number of hospitals: 4,772
Average age of patients: 78.3

<u>For testing of measure exclusions (Section 2b3)</u>
**Dataset 1** (2015 public reporting cohort version 4.0): Medicare Part A Inpatient Claims and Medicare Enrollment Database
Dates of Data: July 1, 2013 – June 30, 2014
Number of index admissions: 6,843,808

Number of hospitals: 4,772
Average age of patients: 78.3

<u>For testing of measure risk adjustment (Section 2b4)</u>
**Dataset 2**: Medicare Part A claims data for calendar years 2007 and 2008 were combined and then randomly split into two equal subsets (development sample and validation sample). Risk

variable selection was done using the development sample, the risk models for each of the five specialty cohorts in the measure were applied to the validation sample and the models' performance was compared. In addition, we re-tested the models in Medicare Part A claims data from calendar year 2009 to look for temporal stability in the models' performance. The number of measured entities and index admissions are listed below by specialty cohort.

Medicine model:
Development sample: 3,085,962 admissions to 4,954 hospitals
Validation sample: 3,082,357 admissions to 4,946 hospitals
2009 sample: 3,032, 518 admissions to 4,908 hospitals
Surgery/gynecology model:
Development sample: 2, 208753 admissions to 4,354 hospitals
Validation sample: 2,208,482 admissions to 4,353 hospitals
2009 sample: 2,109,292 admissions to 4,232 hospitals
Cardiorespiratory model:
Development sample: 1,396562 admissions to 4,810 hospitals
Validation sample: 1,396,855 admissions to 4,806 hospitals
2009 sample: 1,331,539 admissions to 4,718 hospitals
Cardiovascular model:
Development sample: 860,485 admissions to 4,702 hospitals
Validation sample: 861,925 admissions to 4,703 hospitals
2009 sample: 809,520 admissions to 4,641 hospitals
Neurology model:
Development sample: 461,225 admissions to 4,699 hospitals
Validation sample: 461,262 admissions to 4,686 hospitals
2009 sample: 452,743 admissions to 4,609 hospitals

For testing to identify meaningful differences in performance (Section 2b5)
**Dataset 1**

For testing of socioeconomic status (SES) factors and race in risk models (Section 2b4.3)
**Dataset 1** and **Dataset 3**: The American Community Survey (2008-2012)

We examined disparities in performance according to the proportion of patients in each hospital who were of African American race and the proportion who were dual-eligible for both Medicare and Medicaid insurances. We also used the AHRQ SES index score to study the association between performance measures and socioeconomic status.

*Data Elements*
- African American race and dual-eligible status (i.e., enrolled in both Medicare and Medicaid) patient-level data are obtained from CMS enrollment data (**Dataset 1**)
- Validated AHRQ SES index score is a composite of 7 different variables found in the census data (**Dataset 3**)

**Data for the ACR measure**

For reliability testing (Section 2a2)
**Dataset 6:** For reliability testing, we performed stratified random sampling with the **2015 Data.** We randomly split this dataset into two samples; half of each ACO's assigned/aligned beneficiaries were assigned to **Subset 1** and the other half were assigned to **Subset 2**.

**Dataset 6:** We used the Medicare Part A and Part B claims for the 2015 assigned/aligned SSP and Pioneer beneficiaries (**2015 Data**, reported in 2016) for the majority of the analyses. The differences are detailed below.

**2015 Data:** 416 ACOs
Dates of data: Jan 1, 2015- Dec 31, 2015
Mean age: 78.37
Index admissions: 1,721,598

**Subset 1:** 416 ACOs
Dates of data: Jan 1, 2015- Dec 31, 2015
Index admissions: 860,149
Average age of patients: 78.4

**Subset 2:** 416 ACOs
Dates of data: Jan 1, 2015- Dec 31, 2015
Index admissions: 861,449
Average age of patients: 78.5

For testing of measure exclusions (Section 2b3)
**Dataset 6:** We used the Medicare Part A and Part B claims for the 2015 assigned/aligned SSP and Pioneer beneficiaries (**2015 Data**, reported in 2016).

For validity testing (Section 2b2)
**Dataset 5 and Dataset 6:** We assessed measure validity empirically using the **2014 scoring and 2015 scoring datasets**. The scoring datasets include calculated measure rates and performance scores, using the ACO program scoring methodology, for each measure in the ACO measure set. There are three fewer ACOs included in the 2014 scoring data set and six fewer ACOs included in the 2015 scoring data set due to program terminations. The correlation analysis for ACO reported measures (i.e. Web Interface measures) included only ACOs that completely reported and excluded 2015 starters (ACOs receive full points in their first performance year) (n=313).

**Dataset 5:** The 2014 Data includes Medicare Part A and Part B claims for the 2014 assigned/aligned SSP and Pioneer beneficiaries (reported in 2015)
2014 Data: 360 ACOs
Dates of data: Jan 1, 2014- Dec 31, 2014
Index admissions: 1,311,746
Average age of patients: 78.61

**2014 Scoring Dataset:** 357 ACOs
Dates of data: Jan 1, 2014 - Dec 31, 2014

Index admissions: 1,311,642

**2015 Scoring Dataset** (From **Dataset 6**)**:** 410 ACOs
Dates of data: Jan 1, 2015 - Dec 31, 2015
Number of index admissions: 1,678,654
Average age of patients: 78.3

For testing of measure risk adjustment (Section 2b4)
**Dataset 5 and Dataset 6:** We combined the Medicare Part A and Part B claims for the 2014 and 2015 assigned/aligned SSP and Pioneer and then randomly split into two equal subsets (**Development Sample** and **Validation Sample**). The number of measured entities and index admissions are listed below by specialty cohort.
Medicine model:
Development Sample: 639,278 admissions for 449 ACOs
Validation Sample: 639,625 admissions for 449 ACOs
Surgical Cohort:
Development Sample: 383,190 admissions for 449 ACOs
Validation Sample: 382,190 admissions for 448 ACOs
Cardiorespiratory Cohort:
Development Sample: 249,649 admissions for 449 ACOs
Validation Sample: 249,992 admissions for 449 ACOs
Cardiovascular Cohort:
Development Sample: 151,907 admissions for 449 ACOs
Validation Sample: 151,790 admissions for 448 ACOs
Neurology Cohort
Development Sample: 92,759 admissions for 448 ACOs
Validation Sample: 92,503 admissions for 448 ACOs

For testing to identify meaningful differences in performance (Section 2b5)
**Dataset 4, Dataset 5, and Dataset 6: 2013 Data, 2014 Data,** and **2015 Data**

**Dataset 4:** The **2013 Data** includes Medicare Part A and Part B claims for the 2013 assigned/aligned SSP and Pioneer beneficiaries (reported in 2014)
**2013 Data:** 243 ACOs
Dates of data: Jan 1, 2013- Dec 31, 2013
Index admissions: 915,855
Average age of patients: 78.80

For testing of socioeconomic status factors, including dual eligibility, in risk models (Section 2b4.3)
**Dataset 6 and Dataset 7:** The American Community Survey (2010-2014) (**2015 Data** and **SES Dataset)**

We examined disparities in performance based on the proportion of patients in each ACO who were dual-eligible for both Medicare and Medicaid programs. We also examined the association between performance and socioeconomic status, measured by the AHRQ SES index score.

**1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used?** For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Sociodemographic status incorporates socioeconomic variables as well as race into a more concise term. However, given the fact that socioeconomic risk factors are distinct from race and should be interpreted differently, we have decided to keep "socioeconomic status (SES)" and "race" as separate terms.

We selected SES and race variables to analyze after reviewing the literature and examining available national data sources. There is a large body of literature linking various SES factors and African American race to worse health status and higher readmission risk (Blum AB et al., 2014; Eapen ZJ et al. 2015; Gilman M et al., 2014; Hu J et al., 2014; Joynt KE and Jha AK, 2013). Income, education, and occupational level are the most commonly examined variables. However, while literature directly examining how different SES factors or race might influence the likelihood of older, insured, Medicare patients of being readmitted within 30 days of an admission across multiple conditions is more limited, available studies suggest a consistent association between SES/race variables and risk of readmission (Aseltine RH et al., 2015; Gu Q et al., 2014; Arbaje AI et al., 2008). The causal pathways for SES and race variable selection are described below in Section 2b4.3.

For the HWR measure, the SES and race variables used for analysis were:

- Dual eligible status (**Dataset 1**)
- African-American race (**Dataset 1**)
- AHRQ-validated SES index score (percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th-grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room) (**Dataset 3**)

For the ACR measure, the SES variables used for analysis were:

- Dual-eligible status (**Dataset 6**)
- AHRQ-validated SES index score (percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th-grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room) (**Dataset 7**)

In selecting variables, our intent was to be responsive to the NQF guidelines for measure developers in the context of the SDS Trial Period. Our approach has been to examine all patient-level indicators of both SES and race/ethnicity that are reliably available for all Medicare beneficiaries and linkable to claims data and selected those that are most valid.

Previous studies examining the validity of data on patients' race and ethnicity collected by CMS have shown that only the data identifying African American beneficiaries have adequate sensitivity and specificity to be applied broadly in research or measures of quality. While using this variable is not ideal because it groups all non-African American beneficiaries together, it is currently the only race variable available on all beneficiaries across the nation that is linkable to claims data.

We similarly recognize that Medicare-Medicaid dual eligibility has limitations as a proxy for patients' income or assets because it does not provide a range of results and is only a dichotomous outcome. However, the threshold for over 65-year-old Medicare patients is valuable as it takes into account both income and assets which is consistently applied across states. For both our race and the dual-eligible variables, there is a body of literature demonstrating differential health care and health outcomes among beneficiaries indicating that these variables, allow us to examine some of the pathways of interest.

Finally, we selected the AHRQ-validated SES index score because it is a well-validated and widely-used variable that describes the average socioeconomic status of people living in defined geographic areas. Its value as a proxy for patient-level information is dependent on having the most granular level data with respect to communities that patients live in. Currently, the individual data elements used to calculate the score are available at the 5-digit zip code level. The data are not currently available at the 9-digit zip code level.

References:

Arbaje AI, Wolff JL, Yu Q, Powe NR, Anderson GF, Boult C. Post discharge environmental and socioeconomic factors and the likelihood of early hospital readmission among community-dwelling Medicare beneficiaries. The Gerontologist. 2008;48(4):495-504.

Aseltine RH, Jr., Yan J, Gruss CB, Wagner C, Katz M. Connecticut Hospital Readmissions Related to Chest Pain and Heart Failure: Differences by Race, Ethnicity, and Payer. Connecticut medicine. 2015;79(2):69-76.

Blum AB, Egorova NN, Sosunov EA, et al. Impact of socioeconomic status measures on hospital profiling in New York City. Circulation. Cardiovascular quality and outcomes. May 2014; 7(3):391-397.

Eapen ZJ, McCoy LA, Fonarow GC, Yancy CW, Miranda ML, Peterson ED, Califf RM, HernandezAF. Utility of socioeconomic status in predicting 30-day outcomes after heart failure hospitalization. Circ Heart Fail. May 2015; 8(3):473-80.

Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Aff (Millwood). Aug 2014; 33(8):1314-22.

Gu Q, Koenig L, Faerberg J, Steinberg CR, Vaz C, Wheatley MP. The Medicare Hospital Readmissions Reduction Program: potential unintended consequences for hospitals serving vulnerable populations. Health services research. 2014;49(3):818-837.

Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health affairs (Project Hope). 2014; 33(5):778-785.

Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. Jan 23 2013; 309(4):342-3.

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Data Element Reliability

In constructing the measure, we aim to utilize only those data elements from the claims that have both face validity and reliability. We avoid the use of fields that are thought to be coded inconsistently across hospitals or providers. Specifically, we use fields that are consequential for payment and which are audited. We identify such variables through empiric analyses and our understanding of CMS auditing and billing policies and seek to avoid variables which do not meet this standard. For example, "discharge disposition" is a variable in Medicare claims data that is not thought to be a reliable variable for identifying a transfer between two acute care facilities. Thus, we derive a variable using admission and discharge dates as a surrogate for "discharge disposition" to identify hospital admissions involving transfers. This allows us to identify these admissions using variables in the claims data which have greater reliability than the "discharge disposition" variable.

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

Finally, we assess the reliability of the data elements by comparing model variable frequencies and odds ratios from logistic regression models in each new data year.

Measure Score Reliability

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is naturally the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. In line with this thinking, our approach to assessing reliability is to consider the extent to which

assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured once using a random subset of patients, then measured again using a second random subset exclusive of the first, and finally comparing the agreement between the two resulting performance measures across hospitals (Rousson V, et al., 2002).

For test-retest reliability, we randomly sampled half of patients within each hospital in the most recent year of data, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. To the extent that the calculated measures of these two subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement we calculated the intra-class correlation coefficient (ICC) (Shrout P and Fleiss J, 1979), and assessed the values according to conventional standards (Landis and Koch, 1977). Specifically, we used **Dataset 1** split sample and calculated the RSRR for each hospital for each sample. The agreement of the two RSRRs was quantified for hospitals using the intra-class correlation as defined by ICC by Shrout P and Fleiss J (1979).

Using two independent samples provides a stringent estimate of the measure's reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement.

Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less ´signal´, a split sample using a single measurement period would introduce extra noise. This leads to an underestimate in the actual test-retest reliability that would be achieved if the measure were reported using the full measurement period, as evidenced by the Spearman Brown prophecy formula (Spearman CC, 1910; Brown, 1910). We use this to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.

We use the same approach for assessing measure score reliability for the ACR measure.


References:

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

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

Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 2002; 21:3431-3446.

Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979; 86:420-428.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.


**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

**HWR**

The results below are for the HWR measure.

*Data Element Reliability Results*
The frequency of some model variables are assessed in each data year. From year-to-year the frequency of individual variables may increase or decrease slightly. These changes may reflect small changes in rates of comorbidity in the fee-for-service population. For details please see the attached 2015 Measure Updates and Specifications Report. Reports from previous years can be found on QualityNet.

*Measure Score Reliability Results*
There were 6,843,808 admissions in the 2015 public reported measures (**Dataset 1**), with 3,420,728 in one sample and 3,423,080 in the other randomly selected sample. The agreement between the two RSRRs for each hospital was 0.80, which according to the conventional interpretation is "substantial" (Landis J & Koch G, 1977).

**ACR**

The results below are for the ACR measure.

There were 860,149 admissions in **subset 1** and 861,449 admissions in **subset 2** for the 2015 performance year (**Dataset 6**). The intra-class correlation between the two RSRRs among ACO subsets was 0.62, which according to conventional interpretation is "substantial" (Landis J & Koch G, 1977).

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

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i.*e., what do the results mean and what are the norms for the test conducted?*)

For the HWR measure, the ICC score, a form of signal-to-noise analysis, demonstrates substantial agreement across hospital samples.

Similarly, the ICC score for the ACR measure demonstrates substantial agreement across ACO subsets.

**2b2. VALIDITY TESTING**
**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
  ☒ **Empirical validity testing**
  ☒ **Systematic assessment of face validity of performance measure score as an indicator**

of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

**HWR**

*Measure Validity:*

Measure validity is demonstrated through prior validity testing done on our other claims-based measures, through use of established measure development guidelines, and examination of content validity by comparing hospital performance with that on other quality measures.

*Validity of Claims Data:*

Our team has demonstrated for a number of prior measures the validity of claims-based measures for profiling hospitals by comparing either the measure results or individual data elements against the corresponding results and elements from medical records. CMS validated the six NQF-endorsed measures currently in public reporting (acute myocardial infarction [AMI], heart failure, and pneumonia mortality and readmission) with models that used chart-abstracted data for risk-adjustment. Specifically, claims model validation was conducted by building comparable models using abstracted medical chart data for risk adjustment for AMI patients (Cooperative Cardiovascular Project data), (Krumholz HM, et al., 2006) heart failure patients (National Heart Failure data), (Krumholz HM, et al., 2006, Keenan PS, et al., 2006), and pneumonia patients (National Pneumonia Project dataset), (Bratzler DW, et al., 2011). When both models were applied to the same patient population, the hospital risk-standardized rates estimated using the claims-based risk adjustment models had a high level of agreement with the results based on the medical record model, thus supporting the use of the claims-based models for public reporting.

We have also completed two national, multi-site validation efforts for two procedure-based complications measures (for primary elective hip/knee arthroplasty and implantable cardioverter defibrillator [ICD]). Both projects demonstrated strong agreement between complications coded in claims and abstracted medical chart data. These validation efforts suggest that such claims data variables are valid across a variety of conditions.

*Validity Indicated by Established Measure Development Guidelines:*

We developed this measure in consultation with national guidelines for publicly reported outcomes measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcomes measurement set forth in National Quality Forum (NQF) guidance for outcomes measures (National Quality Forum, 2012), CMS Measure Management System guidance, and the guidance articulated in the American Heart Association scientific

statement, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” (Krumholz HM, et al., 2006).

*Validation Against Other Outcomes Measures:*

In order to test the construct validity of the HWR measure, we examined whether hospitals considered “top performers” according to other measures and ranking systems had lower hospital-wide risk-standardized readmission rates than remaining hospitals when applying our measure to the Medicare FFS population. This type of validity testing tests the assumption that hospitals considered top performers have developed an organizational culture of excellence that will manifest itself in better outcomes including lower hospital-wide readmission rates. However, there are multiple challenges associated with this approach:
1. There are many measures and ranking systems available, using a variety of criteria in order to define and select top performers, including: adherence to core processes of care, complications and safety indexes, resource utilization, outcomes, patient satisfaction, and even reputation. “Top performers” on one measure are not the same as “top performers” on another. Moreover, most of these measures are not themselves validated.
2. In many cases, the methodology for identifying “top performers” is proprietary and not transparent.
3. The starting set of hospitals from which different ranking systems select the top performers usually includes only a subset of all acute care hospitals included in the HWR measure; in most cases it is not possible to replicate this starting set exactly.
4. We have not found a ranking system which specifically measures factors most relevant to readmission risk, such as medication reconciliation, patient education, post-discharge follow up, or communication with outpatient clinicians.
After reviewing ranking systems, we selected the following three to use for construct validity testing because they are widely used and their methodology is available to the public:

1. Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey score
http://www.hcahpsonline.org/home.aspx
2. Thomson Reuters 100 top hospitals
http://100tophospitals.com/Portals/2/assets/TOP%2015313%200315%20100%20Top%20Study_web.pdf
3. Joint Commission list of Top Performers on Key Quality Measures
http://www.jointcommission.org/accreditation/top_performers.aspx

**1. HCAHPS**
From the 27 questions in the HCAHPS survey, we selected seven that we felt were most likely to be correlated with readmission rates based on clinical judgment and previously reported results by others (Akamigbo A, 2010, Jha, AK, et al., 2008). Based on previous results we expected to see that patient satisfaction is significantly correlated with hospital readmission rates. For this analysis, we compared 2009 HCAHPS results to 2009 Medicare FFS RSRRs. See results in Section 2b2.3.

**2. Thomson Reuters Top 100 Hospitals**

Given that this measure includes several elements theoretically related to readmission risk, including complications, patient safety, readmissions, and HCAHPS, we felt this measure was a reasonable candidate for construct validity testing. However, since the measure also contains other components such as core measures, expenses, and profitability that would not be expected to correlate with readmission, we expected the analysis to show at best small improvements in readmission performance among top performers. See results in Section 2b2.3.

**3. The Joint Commission's Top Performers on Key Quality Measures program**
Of the Joint Commission's list of 405 top performers, we selected only those 158 hospitals with superior performance in *all four* adult measure sets (HWR is for patients 18 years and older), on the assumption that these hospitals demonstrated hospital-wide performance excellence. We calculated their hospital-wide readmission rates and compared them to those of other hospitals. However, since numerous studies have shown that there is little relationship between performance on core process measures and outcomes including mortality and readmission rates we expected the Joint Commission's top performers to have similar risk-standardized readmission rates as other hospitals, (Bradley E H, et al., 2006; Werner R, et al., 2006; Fonarow GC, et al. 2007; Fonarow GC and Peterson E, 2009; Jha AK, et al., 2009; Patterson M, et al., 2010; Shwartz, M et al., 2011). See results in Section 2b2.3.

**International Classification of Diseases, Ninth Revision (ICD-9) to International Classification of Diseases, Tenth Revision (ICD-10) Conversion**

Statement of Intent
[X] Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.
[ ] Goal was to take advantage of the more specific code set to form a new version of the measure, but fully consistent with the original intent.
[ ] The intent of the measure has changed.

Process of Conversion
ICD-10 codes were initially identified using 2015 General Equivalence Mappings (GEM) software. We then enlisted the help of clinicians with expertise in relevant areas to select and evaluate which ICD-10 codes map to the ICD-9 codes currently in use for this measure. An ICD-9 to ICD-10 crosswalk is attached in field S.2b. (Data Dictionary or Code Table).

We have also examined the updated ICD-9 Map to AHRQ Clinical Classification Software (CCS) crosswalk to the ICD-10 CCS map provided by AHRQ in preparation for the inclusion of ICD 10 data in this measure. Please refer to the ICD-10 CCS map on the AHRQ website.

**ACR**

*Validity Against Other Outcomes Measures:*

To assess the construct validity of the measure at the ACO-level, we examined Pearson correlation statistics between the ACR and other NQF-endorsed measures in the ACO measure set. This type of validity testing assesses whether ACOs with better ACR performance also

perform well on other quality measures with similar constructs. We tested three types of construct validity. First, we assessed the predictive validity of the ACR measure, which is the degree to which the ACR performance in a given performance year correlates with the ACOs' performance on related measures in the subsequent year, using the **2014 Scoring and the 2015 Scoring datasets** for the ACO measures**.** Using two years of data eliminates concern that overlapping data across the related measures could artificially increase the correlations. Next, we assessed convergent validity, the degree to which ACR performance correlates with similar measures and domains in the same performance year, using the **2015 Scoring dataset.** Lastly, we assessed discriminative validity by examining ACR performance correlation with theoretically dissimilar ACO measures in the same performance year using the **2015 Scoring dataset.**

**1. Predictive Validity: Comparison with other NQF endorsed ACO measures in subsequent performance years:**

We assessed the degree to which the ACOs' overall and care coordination/ patient safety performance scores, calculated according to the ACO scoring methodology, correlate with the ACOs' ACR performance. We assessed the association between ACOs' ACR RSRR and their risk-adjusted all-cause admission rate (RSAAR) for the all-cause unplanned admission for patients with diabetes, all-cause unplanned admissions measure for patients with heart failure, and the all-cause unplanned admissions measures for patients with multiple chronic conditions. We expect ACOs with lower RSRRs (better performance) would also exhibit lower RSAARs for unplanned admissions measures (n=321 ACOs).

**2. Convergent Validity: Comparison with related measures in the same performance year Care Coordination/ Patient Safety domain points**

This score was calculated according to the 2015 ACO scoring methodology where similar measures are grouped into domains and are given points according to the benchmarks. These measures include: the ambulatory sensitive conditions admissions for chronic obstructive pulmonary disease or asthma in older adults (PQI #5) and heart failure (PQI #8), screening for future falls (ACO-13), documentation of current medications in the medical record (ACO-39), and the percent of primary care providers who successfully meet meaningful use requirements (ACO-11). Points received for the ACR measure were subtracted from the domain score for this analysis. ACOs that started in 2015 or that did not completely report were excluded from this analysis. ACOs starting in 2015 (n=89 ACOs) were scored based on reporting only and ACOs that did not completely report (n=8 ACOs) have suppressed non-claims-based measure scores as a result of not reporting and were excluded from this analysis. We expected that ACOs with lower RSRRs (better performance) would also perform well (receive more overall points) on the domain (n=313 ACOs).

**Overall Performance Scores**

The overall performance score is the ACO's final performance score, calculated according to the 2015 ACO scoring methodology. This score includes quality improvement points and pay-for-reporting measures. ACOs that started in 2015 (n=89 ACOs) and ACOs that did not completely report (n=8 ACOs) were excluded from this calculation. We expected that ACOs with lower RSRRs would also perform well on the overall performance score (n=313 ACOs).

**Other NQF Endorsed Measures**
We also calculated the correlations for the ACR and the all-cause unplanned admissions for patients with diabetes, heart failure, and multiple chronic conditions. All ACO's were included in this analysis since claims-based measures are not impacted by reporting status (n=410 ACOs).

**4. Discriminative Validity: Comparison with Dissimilar Measures**
**Preventive Care Domain Points**
We compared ACO's performance on the ACR measure with performance on measures in an unrelated domain, the preventive care domain which includes Breast Cancer Screening, Colorectal Cancer Screening, Preventive Care and Screening: Influenza Immunization, BMI Screening, Tobacco Use: Screening and Cessation, High Blood Pressure Screening and Follow-up, and Screening for Clinical Depression and Follow-up Plan. We excluded the Pneumonia Vaccination Status for Older Adults from this analysis as this measure may be associated with reduced admissions and subsequent readmissions. We would not expect the ACR to have a strong association with the preventive care domain measures (n=313 ACOs).

**References:**


Akamigbo A. The Relationship Between Hospital Readmissions and HCAHPS Scores 2010; https://cahps.ahrq.gov/news-and-events/events/UGM12/CAHPS/AkamigboA.pdf

Bradley EH, Herrin J, Elbel, B, McNamara, RL, Magid, DJ, Nallamothu, BK, Krumholz, HM. Hospital quality for acute myocardial infarction: Correlation among process measures and relationship with short-term mortality. Journal of the American Medical Association 2006;296:72-78.

Bratzler DW, Normand SL, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. PLoS One 2011;6(4):e17401.

Fonarow GC et al. Association between performance measures and clinical outcomes for patients hospitalized with heart failure. Journal of the American Medical Association 2007; 297(1):61-70.

Fonarow GC and Peterson E. Heart failure performance measures and outcomes. Journal of the American Medical Association 2009; 302(7): 792-794.

Jha AK, Orav EJ, Zheng J, and Epstein AM. Patients' perception of hospital care in the United States. New England Journal of Medicine 2008;359 (18):1921-31.

Jha AK, Orav EJ, Zheng J, and Epstein AM. Public reporting of discharge planning and rates of readmissions. New England Journal of Medicine 2009; 361(27):2637-45.

Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation 2008;1(1):29-37.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation 2006;113(13):1683-92.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. Circulation 2006;113:1693-1701.

Krumholz HM, Brindis RG, Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. January 24, 2006 2006;113(3):456-462.

National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report REVISED Hospital-Wide Readmission NQF Application January 5, 2012 31

Patterson M, et al. Process of care performance measures and long-term outcomes in patients hospitalized with heart failure. Medical Care 2010; 48(3):210-216.

Shwartz M, et al. How well can we identify the high-performing hospital? Medical Care Research and Review 68;3 (2011):290-310

Werner R and Bradlow E. Relationship between Medicare's hospital compare performance measures and mortality rates. Journal of the American Medical Association 2006;296(22):2694-270.

### 2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

**The following results are for the HWR measure.**

Content validity results from the analyses described above, are presented describing comparisons of hospital performance on the HWR measure and other selected quality metrics.

**1. HCAHPS**

Table 1. shows the correlation (Pearson correlation coefficient) between RSRR and the proportion of patients who responded in that given manner to the question. The analysis includes the 3,723 hospitals that publicly report HCAHPS results.

**Table 1. Correlation between RSRR (2009 Medicare FFS data) and HCAHPS response (N=3,723 hospitals)**

| HCAHPS Question | Correlation |
|---|---|
| Pain was 'sometimes' or 'never' well controlled | 0.34 |
| Patients 'sometimes' or 'never' received help as soon as they wanted | 0.34 |
| Nurses 'sometimes' or 'never' communicated well | 0.33 |
| 'NO' patients would not recommend the hospital | 0.32 |
| Patients were 'sometimes' or 'never' given information about what to do during their recovery at home | 0.32 |
| Patients who gave a rating of '6' or lower | 0.31 |
| Doctors 'sometimes' or 'never' communicated well | 0.21 |

p value for all correlations <0.001REVISED Hospital-Wide Readmission NQF Application January 5, 2012 32

**2. Thomson Reuters Top 100 Hospitals**

Table 2. shows the RSRRs distribution for the top performers in comparison to the rest of hospitals.

**Table 2. Distribution of RSRRs (2009 Medicare FFS data) for the Thomson Reuters Top 100 Hospitals vs. others**

| | On List | Not On List |
|---|---|---|
| Number | 100 | 3017 |
| Mean (SD) | 16.19 (1.39) | 16.65 (1.28) |
| Minimum | 13.77 | 12.51 |
| Lower Quartile | 15.05 | 15.79 |
| Median | 16.06 | 16.51 |
| Upper Quartile | 16.99 | 17.35 |
| Maximum | 19.81 | 22.69 |

**3. The Joint Commission's Top Performers on Key Quality Measures program**

Table 3. shows the distribution of risk-standardized readmission rates of the 158 top performers compared to other hospitals.

**Table 3. Distribution of RSRR (2009 Medicare FFS data) for The Joint Commission's Top Performers vs. Others**

| | On List | Not On List |
|---|---|---|
| Number | 158 | 4630 |
| Mean (SD) | 16.66 (0.99) | 16.61 (1.16) |
| Minimum | 14.18 | 12.51 |
| Lower Quartile | 16.01 | 15.87 |
| Median | 16.64 | 16.49 |
| Upper Quartile | 17.17 | 17.21 |
| Maximum | 19.91 | 22.69 |

**The following results are for the ACR measure.**

The **Predictive Validity** table displays the correlation of the 2014 RSRRs with related NQF-endorsed measures and the overall performance scores.

**Predictive Validity**

| Measure/ Domain | Correlation |
|---|---|
| ACO-36: Diabetes Admissions | 0.5810 |
| ACO-37: Heart Failure Admissions | 0.5537 |
| ACO-38: Multiple Chronic Conditions | 0.5684 |
| ACO-8: ACR | 0.7230 |
| Overall Performance Score | -0.2368 |

All p values are <.001

The **Convergent Validity** table displays the results for the correlations of the RSRRs with similar measures and domains in the 2015 performance year.

**Convergent Validity**

| Measure/ Domain | Correlation |
|---|---|
| Care Coordination/ Patient Safety pay-for-performance points | -0.5197 |
| Overall Performance Score | -0.3101 |
| ACO-36: Diabetes Admissions | 0.6673 |
| ACO-37: Heart Failure Admissions | 0.6721 |
| ACO-38: Multiple Chronic Conditions | 0.6603 |

All p values are <.001

The **Discriminative Validity** table displays the correlations for the RSRR with unrelated measures in performance year 2015.

**Discriminative Validity**

| Measure/ Domain | Correlation |
|---|---|
| Preventive Care Domain points | 0.2094* |
| ACO-14: Influenza Immunization | -0.1087 |
| ACO-16: BMI Screening and Follow-up | 0.0772 |
| ACO-17: Tobacco Use: Screening and Cessation Intervention | -0.1870** |
| ACO-18: Screening for Clinical Depression and Follow-up Plan | -0.0561 |
| ACO-20: Breast Cancer Screening | 0.1874** |
| ACO-19: Colorectal Cancer Screening | -0.2567 ** |
| ACO 21: Screening for High Blood Pressure and Follow-up Documented | 0.0666 |

| | |
|---|---|
| ACO-average SES index among all beneficiaries | -0.014 |

Note: *p<.05 **p<.001

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i.*e., what do the results mean and what are the norms for the test conducted?*)

**HWR**

For the HWR measure, the construct validity analyses demonstrated results consistent with our expectations. There is a significant correlation between patient satisfaction and RSRR as measured by the HWR measure. "Top performers" as defined by Thomson Reuters have lower RSRRs as measured by the HWR measure. On the other hand, hospitals identified by The Joint Commission as having superior performance on all four categories of clinical process measures have identical performance as those with lower performance, consistent with published studies.

**ACR**

For the ACR measure, the construct validity analyses demonstrated results consistent with our expectations. Conventional standards consider a Pearson correlation of 0.37 or larger to be a large association (Cohen, 1988 and 1992). As expected, the ACO ACR performance in 2015 is strongly negatively correlated with the Care Coordination/Patient Safety Domain (higher domain score indicates higher performance) and strongly correlated with three admissions measures. ACR ACO performance in 2014 is strongly correlated with performance in 2015 for several related claims-based measures as well as the ACR ACO performance in 2015. The ACO ACR performance in 2014 is negatively correlated, as expected (higher overall quality score is better), with overall quality performance in 2015. The ACO ACR performance in 2015 is not strongly correlated with other ACO performance on measures that would not be expected to be correlated with the ACR measure. The small correlations are expected, to the extent that high performance on the preventive care domain might influence hospital readmissions. Quality across even unrelated clinical area or processes of care might be weakly correlated because they all reflect the global quality of the measured ACO.

1. Cohen J, Statistical Power Analysis for the Behavioral Sciences, 2nd ed., 1988.
2. Cohen J, A power primer, Psychol Bull, 1992;112(1):155-159.

**2b3. EXCLUSIONS ANALYSIS**

**NA ☐ no exclusions — *skip to section 2b4***

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions and to ensure accurate calculation of the measure. To ascertain impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion (**Dataset 1** and **Dataset 6**). These exclusions are consistent with similar NQF-endorsed outcome measures. Rationales for the exclusions are detailed in data field S.10 (Denominator Exclusions).

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**The following results are for the HWR measure.**

In Dataset 1 (2015 Public Reporting Cohort):

| **Exclusion** | **N** | **%** | **Distribution across hospitals (N=4,802): min, 25th, 50th, 75th, max** |
|---|---|---|---|
| Admitted to PPS-Exempt Cancer Hospitals | 19823 | 0.28 | (0.00, 0.00, 0.00, 0.00, 100.00) |
| Without 30 Days of Post-Discharge Enrollment | 36640 | 0.52 | (0.00, 2.3, 3.1, 4.0, 22.2) |
| Discharged against medical advice (AMA) | 26665 | 0.38 | (0.00, 0.20, 0.50, 1.00, 21.1) |
| Admitted for Primary Psychiatric Diagnosis | 19691 | 0.28 | (0.00, 0.00, 0.10, 0.40, 100.00) |
| Admitted for Rehabilitation | 7152 | 0.10 | (0.00, 0.00, 0.00, 0.00, 100.00) |
| Admitted for Medical Treatment of Cancer | 152288 | 2.15 | (0.00, 0.60, 1.30, 1.90, 55.00) |

**The following results are for the ACR measure.**

In Dataset 6 (2016 Public Reporting Cohort)

| **Exclusion** | **N** | **%** | **Distribution across ACOs (N=416): min, 25th, 50th, 75th, max** |
|---|---|---|---|
| Admitted to PPS-Exempt Cancer Hospitals | 4058 | 0.22 | (0.00, 0.00, 0.07, 0.26, 2.73) |
| Without 30 Days of Post-Discharge Enrollment | 6062 | 0.34 | (0.00, 0.21, 0.30, 0.42, 1.03) |
| Discharged against medical advice (AMA) | 6313 | 0.35 | (0.00, 0.19, 0.30, 0.48, 1.85) |

| | | | |
|---|---|---|---|
| Admitted for Primary Psychiatric Diagnosis | 4483 | 0.25 | (0.00, 0.13, 0.20, 0.31, 1.74) |
| Admitted for Rehabilitation | 1391 | 0.08 | (0.00, 0.00, 0.00, 0.02, 3.65) |
| Admitted for Medical Treatment of Cancer | 37767 | 2.09 | (0.39, 1.62, 1.93, 2.28, 7.40) |

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.* *Note*: ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

**Exclusions applied to the HWR measure cohort:**

1. Patients admitted to Inpatient Prospective Payment System (IPPS)-exempt cancer hospitals account for 0.28% of all index admissions excluded from the initial cohort. Admissions for treatment of cancer are associated with a very different mortality and readmission risk compared with admissions to other IPPS hospitals for treatment of other diseases. Additionally, outcomes for these admissions do not correlate well with outcomes for other types of admissions. (Patients with cancer who are admitted for other diagnoses or for surgical treatment of their cancer remain in the measure).
2. Patients without at least 30 days post-discharge enrollment in FFS Medicare following discharge account for 0.52% of all index admissions excluded from the initial cohort. This exclusion is needed since the 30-day readmission outcome cannot be assessed in patients who do not maintain enrollment for at least 30 days following discharge.
3. Patients discharged against medical advice (AMA) account for 0.38% of all index admissions excluded from the initial index cohort. This exclusion is needed for acceptability of the measure to hospitals, who do not have the opportunity to adequately deliver full care and prepare the patient for discharge.
4. Patients admitted for primary psychiatric diagnoses account for 0.28% of all index admissions excluded from the initial cohort. This exclusion is needed because these patients are typically cared for in separate psychiatric or rehabilitation centers which are not comparable to acute care hospitals.
5. Patients admitted for rehabilitation account for 0.10% of all index admissions excluded from the initial cohort. This exclusion is needed because patients admitted for rehabilitation are not admitted for treatment of acute illness and the care provided in rehabilitation centers is not comparable to care provided in acute care hospitals.
6. Patients admitted for medical treatment of cancer account for 2.15% of all index admissions excluded from the initial cohort. Admissions for treatment of cancer are associated with a very different mortality and readmission risk compared with admissions to other IPPS hospitals for treatment of other diseases. Additionally, outcomes for these admissions do not correlate well with outcomes for other types of admissions. (Patients with cancer who are admitted for other diagnoses or for surgical treatment of their cancer remain in the measure).

**Exclusions applied to the ACR measure cohort:**

1. Patients admitted to Inpatient Prospective Payment System (IPPS)-exempt cancer hospitals account for 0.22% of all index admissions excluded from the initial cohort.
2. Patients without at least 30 days post-discharge enrollment in Medicare FFS following discharge account for 0.34% of all index admissions excluded from the initial cohort.
3. Patients discharged against medical advice (AMA) account for 0.35% of all index admissions excluded from the initial index cohort.
4. Patients admitted for primary psychiatric diagnoses account for 0.25% of all index admissions excluded from the initial cohort.
5. Patients admitted for rehabilitation account for 0.08% of all index admissions excluded from the initial cohort.
6. Patients admitted for medical treatment of cancer account for 2.09% of all index admissions excluded from the initial cohort.

In general, these exclusions have a minor impact on the performances.

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**

***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.***

**2b4.1. What method of controlling for differences in case mix is used?**

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 33 risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

**2b4.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

Please see data dictionary for coefficients from all cohorts.

**2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

N/A

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

**HWR**

Our approach to risk adjustment was tailored to and appropriate for a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz HM, et al., 2006). The measure estimates hospital-level 30-day all-cause RSRRs using hierarchical logistic regression models. In brief, the approach simultaneously models data at the patient and hospital levels to account for variance in patient outcomes within and between hospitals, (Normand S-LT, Shahian DM, 2007).

At the patient level, it models the log-odds of hospital readmission within 30 days of discharge using age, selected clinical covariates, and a hospital-specific intercept. At the hospital level, the approach models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of a readmission at the hospital, after accounting for patient risk. The hospital-specific intercepts are given a distribution to account for the clustering (non-independence) of patients within the same hospital (Normand S-LT, Shahian DM, 2007). If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

Admissions are assigned to one of five mutually exclusive specialty cohort groups consisting of related conditions or procedures. For each specialty cohort group, the standardized readmission ratio (SRR) is calculated as the ratio of the number of "predicted" readmissions to the number of "expected" readmissions at a given hospital. For each hospital, the numerator of the ratio is the number of readmissions within 30 days predicted based on the hospital's performance with its observed case mix and service mix, and the denominator is the number of readmissions expected based on the nation's performance with that hospital's case mix and service mix. This approach is analogous to a ratio of "observed" to "expected" used in other types of statistical analyses. It conceptually allows a particular hospital's performance, given its case mix and service mix, to be compared to an average hospital's performance with the same case mix and service mix. Thus, a lower ratio indicates lower-than-expected readmission rates or better quality, while a higher ratio indicates higher-than-expected readmission rates or worse quality.

For each specialty cohort, the "predicted" number of readmissions (the numerator) is calculated by using the coefficients estimated by regressing the risk factors (found in the attached Data Dictionary) and the hospital-specific intercept on the risk of readmission. The estimated hospital-specific intercept for each cohort is added to the sum of the estimated regression coefficients multiplied by patient characteristics. The results are transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" number of readmissions (the denominator) is obtained in the same manner, but a common intercept using all hospitals in our sample is added in place of the hospital-specific intercept. The results are transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimate the model coefficients using the data in that period.

The specialty cohort SRRs are then pooled for each hospital using a volume-weighted geometric mean to create a hospital-wide composite SRR. The composite SRR is multiplied by the national observed readmission rate to produce the RSRR.

*Data Source*

The HWR risk-adjustment models use only inpatient claims data (history and current) in order to make it feasible to implement with Medicare data, and to make it applicable to all-payer data, which are typically restricted to inpatient claims.

The HWR measure uses CMS-CCs (Horwitz L, Partovian C, Lin Z, et al. 2012), the grouper used in previous CMS risk-standardized outcomes measures, to group ICD-9-CM codes into comorbid risk adjustment variables, since four CMS condition-specific claims-based readmission models that use this grouper to define variables for risk adjustment have been validated against models that use chart-abstracted data for risk adjustment (Pope G, et al., 2000, Keenan PS, Normand SL, Lin Z, et al., 2008, Krumholz HM, Lin Z, Drye EE, et al. 2011).

*Approach to Variable Selection:*

In order to select the comorbid risk variables, we developed a "starter" set of 30 variables drawn from previous readmission measures (AMI, heart failure, pneumonia, hip and knee arthroplasty, and stroke). Next we reviewed all the remaining CMS-CCs and determined on a clinical basis whether they were likely to be relevant to an all-condition measure. We selected 11 additional risk variables to consider.

Using data from the index admission and any admission in the prior 12 months, we ran a standard logistic regression model for every discharge condition category with the full set of candidate risk adjustment variables. We compared odds ratios for different variables across different condition categories (excluding condition categories with fewer than 700 readmissions due to the number of events per variable constraints). We selected the final set of comorbid risk variables based on the following principles:

• We excluded risk variables that were statistically significant for very few condition categories, given that they would not contribute much to the overall models.

• We excluded risk variables that behaved in clinically incoherent ways. For example, we dropped risk variables that sometimes increased risk and sometimes decreased risk, when we could not identify a clinical rationale for the differences.

• We excluded risk variables that were predominantly protective when we felt this protective effect was not clinically reasonable but more likely reflected coding factors. For example, drug/alcohol abuse without dependence (CC 53) and delirium and encephalopathy (CC 48) were both protective for readmission risk although clinically they should increase patients' severity of illness.

• Where possible, we grouped together risk variables that were clinically coherent and carried similar risks across condition categories. For example, we combined coronary artery disease (CCs 83-84) with cerebrovascular disease (CCs 98, 99, and 103).

• We examined risk variables that had been combined in previous CMS publicly reported measures, and in one instance separated them: for cancers, the previous measures generally pool 5 categories of cancers (CCs 8 to 12), together. In our analysis, lung cancer (CC 8) and other severe cancers (CC 9) carried higher risks, so we separated them into a distinct risk variable and grouped other major cancers (CC 10), benign cancers (CC 11), and cancers of the urinary and GI tracts (CC 12) together. Consistent with other publicly reported measures, we also left metastatic cancer/leukemia (CC 7) as a separate risk variable.

Complications occurring during hospitalization are not comorbid illnesses, may reflect hospital quality of care, and therefore should not be used for risk adjustment. Hence, conditions that may represent adverse outcomes due to care received during the index hospital stay are not included in the risk-adjusted model; see Table 5 in Section 2a1.13. CCs on this list were not counted as a risk variable in our analyses if they appeared only on the index admission.

*Service mix adjustment:*

The measure includes many different discharge condition categories that differ in their baseline readmission risks. In addition, hospitals differ in their relative distribution of these condition categories (service mix). To adjust for service mix, the measure uses an indicator variable for the discharge condition category in addition to risk variables for comorbid conditions. The models include a condition-specific indicator for all condition categories with sufficient volume (defined as those with more than 1,000 admissions nationally in a given year for Medicare FFS data) as well as a single indicator for conditions with insufficient volume in each model.

*Socioeconomic Status (SES) Factors and Race*

SES factors and race for examination were based on a review of literature, conceptual pathways, and feasibility. In Section 1.8, we describe the variables that we considered and analyzed based on this review. Below we describe the pathways by which SES and race may influence 30-day readmission.

Our conceptualization of the pathways by which patient SES or race affects 30-day readmission is informed by the literature.

*SES and Race Variables and HWR*

To examine the relationship between SES and race variables and hospital 30-day, hospital-wide, all-cause, unplanned readmission following hospitalization, a literature search was performed with the following exclusion criteria: international studies, articles published more than 10 years ago, articles without primary data, articles using Veterans Affairs (VA) databases as the primary data source, and articles not explicitly focused on SES or race and readmission across multiple conditions. One hundred and sixty nine articles were initially reviewed, and one hundred and fifty five studies were excluded from full-text review based on the above criteria. Studies indicate that SES/race variables were associated with increased risk of readmission across multiple major illnesses and conditions (Aseltine RH, et al., 2015; Mitchell SE, et al., 2012; Odonkor CA, et al., 2015; Herrin J, et al., 2015; Gu Q, et al., 2014, Kim H, et al., 2010; Kangovi S, et al., 2012; Iloabuchi TC, 2014; Beck AF, et al., 2012; Arbaje AI, et al., 2008; Hu J, 2014; Nagasako EM, et al., 2014; Joynt, KE, et al., 2013), though there may not be a significant effect on hospital-level profiling (Blum AB, et al., 2014).

*SES and Race Variable Selection*

Although some recent literature evaluates the relationship between patient SES or race and the readmission outcome, few studies directly address causal pathways or examine the role of the hospital in these pathways. Moreover, the current literature examines a wide range of conditions and risk variables with no clear consensus on which risk factors demonstrate the strongest relationship with readmission. The SES factors that have been examined in the readmission literature can be categorized into three domains: (1) patient-level variables, (2) neighborhood/community-level variables, and (3) hospital-level variables. Patient-level variables

describe characteristics of individual patients, and range from the self-reported or documented race or ethnicity of the patient to the patient's income or education level (Eapen ZJ, et al., 2015; Hu J, et al., 2014). Neighborhood/community-level variables use information from sources such as the American Community Survey (ACS) as either a proxy for individual patient-level data or to measure environmental factors. Studies using these variables use one dimensional measures such as median household income or composite measures such as the Agency for Healthcare Research and Quality (AHRQ)-validated SES index score (Blum AB, et al., 2014). Hospital-level variables measure attributes of the hospital which may be related to patient risk. Examples of hospital-level variables used in studies are ZIP code characteristics aggregated to the hospital level or the proportion of Medicaid patients served in the hospital (Gilman M, et al., 2014; Joynt KE and Jha AK, 2013).

The conceptual relationship, or potential causal pathways by which these possible SES risk factors and race/ ethnicity influence the risk of readmission following an acute illness or major surgery, like the factors themselves, are varied and complex. There are at least four potential pathways that are important to consider.

1. Relationship of socioeconomic status (SES) factors or race to health at admission. Patients who have lower income/education/literacy or unstable housing may have a worse general health status and may present for their hospitalization or procedure with a greater severity of underlying illness. These SES risk factors, which are characterized by patient-level or neighborhood/community-level (as proxy for patient-level) variables, may contribute to worse health status at admission due to competing priorities (restrictions based on job, lack of childcare), lack of access to care (geographic, cultural, or financial), or lack of health insurance. Given that these risk factors all lead to worse general health status, this causal pathway should be largely accounted for by current clinical risk-adjustment.

In addition to SES risk factors, studies have shown that worse health status is more prevalent among African-American patients compared with white patients. The association between race and worse health is in part mediated by the association between race and SES risk factors such as poverty or disparate access to care associated with poverty or neighborhood. The association is also mediated through bias in healthcare as well as other facets of society.

2. Use of low-quality hospitals. Patients of lower income, lower education, or unstable housing have been shown not to have equitable access to high quality facilities because such facilities are less likely to be found in geographic areas with large populations of poor patients; thus patients with low income are more likely to be seen in lower quality hospitals, which can contribute to increased risk of readmission following hospitalization (Jha AK, et al., 2011; Reames BN, et al., 2014). Similarly African-American patients have been shown to have less access to high quality facilities compared with white patients (Skinner J, et al., 2005).

3. Differential care within a hospital. The third major pathway by which SES factors or race may contribute to readmission risk is that patients may not receive equivalent care within a facility. For example, African-American patients have been shown to experience differential, lower quality, or discriminatory care within a given facility (Trivedi AN, et al., 2014). Alternatively, patients with SES risk factors such as lower education may require differentiated care – e.g. provision of lower literacy information – that they do not receive.

4. Influence of SES on readmission risk outside of hospital quality and health status. Some SES risk factors, such as income or wealth, may affect the likelihood of readmission without directly

affecting health status at admission or the quality of care received during the hospital stay. For instance, while a hospital may make appropriate care decisions and provide tailored care and education, a lower-income patient may have a worse outcome post-discharge due to competing economic priorities or a lack of access to care outside of the hospital.

These proposed pathways are complex to distinguish analytically. They also have different implications on the decision to risk adjust or not. We, therefore, first assessed if there was evidence of a meaningful effect on the risk model to warrant efforts to distinguish among these pathways. Based on this model and the considerations outlined in Section 1.8, the following SES and race variables were considered:

• Dual-eligible status

• African American race

• AHRQ SES index

We assessed the relationship between the SES variables and race with the outcome and examined the incremental effect in a multivariable model. For this measure, we also examined the extent to which the addition of any one of these variables improved model performance or changed hospital results.

One concern with including SES or race factors in a model is that their effect may be at either the patient or the hospital level. For example, low SES may increase the risk of readmission because patients of low SES have an individual higher risk (patient-level effect) or because patients of low SES are more often admitted to hospitals with higher overall readmission rates (hospital-level effect). Thus, as an additional step, we performed a decomposition analysis to assess the independent effects of the SES and race variables at the patient level and the hospital level. If, for example, all the elevated risk of readmission for patients of low SES was due to lower quality/higher readmission risk in hospitals with more patients of low SES, then a significant hospital-level effect would be expected with little-to-no patient-level effect. However, if the increased readmission risk was solely related to higher risk for patients of low SES regardless of hospital effect, then a significant patient-level effect would be expected and a significant hospital-level effect would not be expected.

Specifically, we decomposed each of the SES and race variables as follows: Let Xij be a binary indicator of the SES or race status of the ith patient at the jth hospital, and Xj the percent of patients at hospital j with Xij = 1. Then we rewrote Xij = (Xij- Xj) + Xj □ Xpatient+ Xhospital. The first variable, Xpatient, represents the effect of the risk factor at the patient level (sometimes called the "within" hospital effect), and the second, Xhospital, represents the effect at the hospital level (sometimes called the "between" hospital effect). By including both of these in the same model, we can assess whether these are independent effects, or whether only one of these effects contributes. This analysis allows us to simultaneously estimate the independent effects of: 1) hospitals with higher or lower proportions of low SES patients or African-American patients on the readmission rate of an average patient; and 2) a patient's SES or race on their own readmission rates when seen at an average hospital.

It is very important to note, however, that even in the presence of a significant patient-level effect and absence of a significant hospital-level effect, the increased risk could be partly or entirely due to the quality of care patients receive in the hospital. For example, biased or differential care provided within a hospital to low-income patients as compared to high-income patients would

exert its impact at the level of individual patients, and therefore be a patient-level effect. It is also important to note that the patient-level and hospital-level coefficients cannot be quantitatively compared because the patient's SES circumstance or race in the model is binary whereas the hospitals' proportion of low SES patients or African-American patients is continuous.

**ACR**

In considering modifying this measure for the ACO program, we were guided by a conceptual framework outlining the relationships between potential clinical and contextual factors and rates of readmission at the ACO level. Importantly, many factors other than traditional medical care delivered in the office or hospital settings will impact the likelihood of readmission. For example, ACO's practicing in communities where patients have limited access to transportation, healthy foods and recreational facilities, may have less success in promoting healthy behaviors among patients; this may in-turn impact readmission rates. Recognition of and attention to the health environment may be important for achieving the goals of better care, better health, and lower costs and thus, shared savings.

Our conceptual model recognizes patient-level demographic and clinical factors, along with four contextual domains that may influence ACO performance: (1) Physical environment (e.g., green spaces, safe streets); (2) Community resources (e.g., home health, senior services); (3) Patient resources (e.g., social support, transportation, income); and (4) Patient behavior/personal preferences (e.g., exercise, diet, advanced care directives, preference for intervention).

The model also recognizes the capacity of ACOs to mitigate the effects of many contextual factors on rates of admissions, encompassing both SES and non-SES variables. Adjusting for contextual factors would obscure important differences in ACO quality and could serve as a disincentive for ACO's to engage with such factors. ACO's should and do influence a broad range of patient and community level factors that can mitigate the risk of readmission associated with the contextual environment.

We did, however, conduct analyses of SES factors to further inform the committee's deliberation (see 2b4.4b). To examine the influence of community level contextual factors, we utilize a patient-level variable, the AHRQ SES index, that is validated as a measure of community level contextual factors. We also examined the influence of dual Medicare and Medicaid eligibility status on ACR measure performance.

**References:**

Arbaje AI, Wolff JL, Yu Q, Powe NR, Anderson GF, Boult C. Postdischarge environmental and socioeconomic factors and the likelihood of early hospital readmission among community-dwelling Medicare beneficiaries. The Gerontologist. 2008;48(4):495-504

Aseltine RH, Jr., Yan J, Gruss CB, Wagner C, Katz M. Connecticut Hospital Readmissions Related to Chest Pain and Heart Failure: Differences by Race, Ethnicity, and Payer. Connecticut medicine. 2015;79(2):69-76.

Beck AF, Simmons JM, Huang B, Kahn RS. Geomedicine: area-based socioeconomic measures for assessing risk of hospital reutilization among children admitted for asthma. American journal of public health. 2012;102(12):2308-2314.

Blum AB, NN Egorova, E. A. Sosunov, A. C. Gelijns, E. DuPree, A. J. Moskowitz, A. D. Federman, D. D. Ascheim and S. Keyhani. "Impact of Socioeconomic Status Measures on Hospital Profiling in New York City." Circ Cardiovasc Qual Outcomes 7, no. 3 (2014): 391-7.

Eapen ZJ, McCoy LA, Fonarow GC, Yancy CW, Miranda ML, Peterson ED, Califf RM, Hernandez AF. Utility of socioeconomic status in predicting 30-day outcomes after heart failure hospitalization. Circ Heart Fail. May 2015; 8(3):473-80.

Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Aff (Millwood). Aug 2014; 33(8):1314-22.

Gu Q, Koenig L, Faerberg J, Steinberg CR, Vaz C, Wheatley MP. The Medicare Hospital Readmissions Reduction Program: potential unintended consequences for hospitals serving vulnerable populations. Health services research. 2014;49(3):818-837.

Herrin J, St Andre J, Kenward K, Joshi MS, Audet AM, Hines SC. Community factors and hospital readmission rates. Health services research. 2015;50(1):20-39.

Horwitz L, Partovian C, Lin Z, et al. Hospital-Wide All-Cause Unplanned Readmission Measure: Final Technical Report. 2012;

https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1228772504318

Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health affairs (Project Hope). 2014; 33(5):778-785.

Iloabuchi TC, Mi D, Tu W, Counsell SR. Risk factors for early hospital readmission in low-income elderly adults. Journal of the American Geriatrics Society. 2014;62(3):489-494.

Jha AK, Orav EJ, Epstein AM. Low-quality, high-cost hospitals, mainly in South, care for sharply higher shares of elderly black, Hispanic, and Medicaid patients. Health Affairs 2011; 30:1904-11.

Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. Jan 23 2013; 309(4):342-3.

Joynt, K. E., E. J. Orav and A. K. Jha. "Thirty-Day Readmission Rates for Medicare Beneficiaries by Race and Site of Care." JAMA 305, no. 7 (2011): 675-81.

Kangovi S, Grande D, Meehan P, Mitra N, Shannon R, Long JA. Perceptions of readmitted patients on the transition from hospital to home. Journal of hospital medicine. 2012;7(9):709-712.

Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation 2008;1(1):29-37.

Kim H, Ross JS, Melkus GD, Zhao Z, Boockvar K. Scheduled and unscheduled hospital readmissions among patients with diabetes. The American journal of managed care. 2010;16(10):760-767.

Krumholz HM, Brindis RG, Brush JE, et al. 2006. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement

From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation 113: 456-462.

Krumholz HM, Lin Z, Drye EE, et al. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Circulation 2011;4(2):243-52.

Mitchell SE, Sadikova E, Jack BW, Paasche-Orlow MK. Health literacy and 30-day postdischarge hospital utilization. Journal of health communication. 2012;17 Suppl 3:325-338.Nagasako EM, Reidhead M, Waterman B, Dunagan WC. Adding socioeconomic data to hospital readmissions calculations may produce more useful results. Health affairs (Project Hope). 2014;33(5):786-791

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.

Odonkor CA, Hurst PV, Kondo N, Makary MA, Pronovost PJ. Beyond the Hospital Gates: Elucidating the Interactive Association of Social Support, Depressive Symptoms, and Physical Function with 30-Day Readmissions. American journal of physical medicine & rehabilitation / Association of Academic Physiatrists. 2015;94(7):555-567

Pope, G., et al., Principal Inpatient Diagnostic Cost Group Models for Medicare Risk Adjustment. Health Care Financing Review, 2000. 21(3):26.

Reames BN, Birkmeyer NJ, Dimick JB, Ghaferi AA. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. JAMA surgery 2014; 149:475-81.

Skinner J, Chandra A, Staiger D, Lee J, McClellan M. Mortality after acute myocardial infarction in hospitals that disproportionately treat black patients. Circulation 2005; 112:2634-41.

Trivedi AN, Nsa W, Hausmann LR, et al. Quality and equity of care in U.S. hospitals. The New England journal of medicine 2014; 371:2298-308.

**2b4.4a. What were the statistical results of the analyses used to select risk factors?**

The final variables for each of the five risk models with associated odds ratios (**Dataset 1**) are shown in the attached Data Dictionary or Code Table 2.2b.

**2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)**

**HWR**

*Variation in prevalence of the factor across measured entities*
The prevalence of SDS factors in the HWR cohort varies across measured entities. The median percentage of Medicaid patients is 14.9% (interquartile range [IQR] 9.8%-22.6 %). The median

percentage of African-American patients is 2.2% (IQR 0.0%-9.4%). The median percentage of low SES AHRQ indicator patients is 19.4% (IQR 5.0%-57.3%).

*Empirical association with the outcome (bivariate)*
The patient-level observed hospital wide readmission rate is higher for Medicaid patients, 19.3%, compared with 14.8% for all other patients. The readmission rate for African-American patients was also higher at 19.2% compared with 15.1% for patients of all other races. Similarly the readmission rate for patients in the lowest SES quartile by AHRQ Index was 16.8% compared with 15.1% for all other patients.

*Incremental effect of SDS variables in a multivariable model*
We then examined the strength and significance of the SDS variables in the context of a multivariable model. Consistent with the above findings, when we include any of these variables in a multivariate model that includes all of the claims-based clinical variables the effect size of each of these variables is small. We also find that the c-statistic is essentially unchanged with the addition of any of these variables into the model. Furthermore we find that the addition of any of these variables into the model has little to no effect on hospital performance. We examined the change in hospitals' RSRRs with the addition of any of these variables. The mean median absolute change in hospitals' RSRRs when adding a Medicaid indicator is 0.004% (interquartile range [IQR] -0.017% – 0.024%, minimum -0.309% – maximum 0.135%) with a correlation coefficient between RSRRs for each hospital with and without Medicaid added of 0.998. The median absolute change in hospitals' RSRRs when adding a race indicator is 0.011% (IQR -0.010% – 0.033%, minimum -0.671% – maximum 0.130%) with a correlation coefficient between RSRRs for each hospital with and without race added of 0.998. The median absolute change in hospitals' RSRRs when adding a low SES AHRQ indicator is 0.007% (IQR -0.033% – 0.036%, minimum -0.322% – maximum 0.135%) with a correlation coefficient between RSRRs for each hospital with and without low SES added of 0.997.

As an additional step, a decomposition analysis was performed. The results are described in the table below.

The patient-level and hospital-level dual-eligible, race, and low AHRQ SES Index effects were significantly associated with each of the hospital wide readmission models (Medicine, Surgery, Cardiorespiratory, Cardiovascular, and Neurology) in the decomposition analysis. If the dual-eligible, race, or low AHRQ SES Index variables are used to adjust for patient-level differences, then some of the differences between hospitals would also be adjusted for, potentially obscuring a signal of hospital quality.

Given these findings and the complex pathways that could explain any relationship between SES or race with readmission, we did not incorporate SES variables or race into the measure.

## HWR Decomposition Analysis

| Parameter | Estimate (Standard Error) | P-value |
|---|---|---|
| **Dual Eligible – Patient-Level – Medicine** | 0.0599 (0.00433) | <.0001 |
| **Dual Eligible – Hospital-Level – Medicine** | 0.3207 (0.0177) | <.0001 |
| **Dual Eligible – Patient-Level – Surgery** | 0.1483 (0.00794) | <.0001 |
| **Dual Eligible – Hospital-Level – Surgery** | 0.4743 (0.0332) | <.0001 |
| **Dual Eligible – Patient-Level – Cardio Respiratory** | 0.1043 (0.00634) | <.0001 |
| **Dual Eligible – Hospital-Level – Cardio Respiratory** | 0.4148 (0.0269) | <.0001 |
| **Dual Eligible – Patient-Level – Cardiovascular** | 0.1607 (0.0101) | <.0001 |
| **Dual Eligible – Hospital-Level – Cardiovascular** | 0.5318 (0.0418) | <.0001 |
| **Dual Eligible – Patient-Level – Neurology** | 0.0874 (0.0129) | <.0001 |
| **Dual Eligible – Hospital-Level – Neurology** | 0.4997 (0.0526) | <.0001 |
| **African American – Patient-Level – Medicine** | 0.0374 (0.00558) | <.0001 |
| **African American – Hospital-Level – Medicine** | 0.3208 (0.0119) | <.0001 |
| **African American – Patient-Level – Surgery** | 0.0959 (0.0103) | <.0001 |
| **African American – Hospital-Level – Surgery** | 0.4423 (0.0214) | <.0001 |
| **African American – Patient-Level – Cardio Respiratory** | 0.0470 (0.00884) | <.0001 |
| **African American – Hospital-Level – Cardio Respiratory** | 0.3386 (0.0186) | <.0001 |
| **African American – Patient-Level – Cardiovascular** | 0.0763 (0.0131) | <.0001 |
| **African American – Hospital-Level – Cardiovascular** | 0.3501 (0.0269) | <.0001 |
| **African American – Patient-Level – Neurology** | 0.1200 (0.0155) | <.0001 |
| **African American – Hospital-Level – Neurology** | 0.5252 (0.0331) | <.0001 |
| **AHRQ SES Index – Patient-Level – Medicine** | 0.0249 (0.00444) | <.0001 |
| **AHRQ SES Index – Hospital-Level – Medicine** | 0.0788 (0.00653) | <.0001 |
| **AHRQ SES Index – Patient-Level – Surgery** | 0.0349 (0.00689) | <.0001 |

| | | |
|---|---|---|
| **AHRQ SES Index – Hospital-Level – Surgery** | 0.1254 (0.0120) | <.0001 |
| **AHRQ SES Index – Patient-Level – Cardio Respiratory** | 0.0376 (0.00661) | <.0001 |
| **AHRQ SES Index – Hospital-Level – Cardio Respiratory** | 0.1105 (0.00910) | <.0001 |
| **AHRQ SES Index – Patient-Level – Cardiovascular** | 0.0307 (0.00943) | 0.0011 |
| **AHRQ SES Index – Hospital-Level – Cardiovascular** | 0.1375 (0.0149) | <.0001 |
| **AHRQ SES Index – Patient-Level – Neurology** | 0.0544 (0.0125) | <.0001 |
| **AHRQ SES Index – Hospital-Level – Neurology** | 0.1314 (0.0198) | <.0001 |

**ACR**

We performed analyses to assess the effect of SES on the ACR measure. These analyses are informative for future measure use, but the decision not to adjust for SES in this measure was based on conceptual factors and not on the results of these statistical analyses (see 2b4.3.).

*Variation in prevalence of the factor across measured entities*
The prevalence of SES factors in the ACR cohort varies across measured entities. The median percentage of dual-eligible patients is 8.0% (interquartile range [IQR] 5.1%-13.3%). The median percentage of low SES AHRQ indicator patients is 48.9% (IQR 31.8%-66.4%).

*Empirical association with the outcome (bivariate)*
The patient-level observed hospital wide readmission rate is higher for dual-eligible patients, 16.0%, compared with 14.7% for all other patients. Similarly, the readmission rate for patients in the lowest SES quartile by AHRQ Index was 15.6% compared with 14.2% for all other patients.

*Incremental effect of SES variables in a multivariable model*
We examined the strength and significance of the SES variables in the context of a multivariable model. When we include any of these variables in a multivariate model that includes all of the claims-based clinical variables the effect size of each of these variables is small. We also find that the c-statistic is essentially unchanged with the addition of any of these variables into the model. Furthermore, we find that the addition of any of these variables into the model has little to no effect on ACO performance. We examined the change in ACOs' RSRRs with the addition of these variables. The median absolute change in ACOs' RSRRs when adding a dual-eligible indicator is -0.105% (interquartile range [IQR] -0.305% – 0.216%, minimum -0.794% – maximum 2.426%) with a correlation coefficient between RSRRs for each ACO with and without dual-eligible added of 0.997. The median absolute change in ACOs' RSRRs when

adding a low SES AHRQ indicator is -0.001% (IQR -0.054% – 0.046%, minimum -0.200% – maximum 0.174%) with a correlation coefficient between RSRRs for each hospital with and without low SES added of 0.9999.

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps—do not just name a method; what statistical analysis was used*)

***Approach to assessing model performance (Dataset 1** and **Dataset 2)***
We tested the performance of the model for **Dataset 1** and **Dataset 2** described in section 1.7. We computed three summary statistics for assessing model performance (Harrell and Shih, 2001) for the development and validation cohort:

***Discrimination Statistics***
(1) Area under the receiver operating characteristic (ROC) curve (the c-statistic) is the probability that predicting the outcome is better than chance, which is a measure of how accurately a statistical model is able to distinguish between a patient with and without an outcome.
(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; therefore, we would hope to see a wide range between the lowest decile and highest decile.)

***Calibration Statistics*** **(Dataset 2)**
(3) Over-fitting indices (over-fitting refers to the phenomenon in which a model accurately describes the relationship between predictive variables and outcome in the development dataset but fails to provide valid predictions in new patients.)

References:
Harrell FE and Shih YCT. Using full probability models to compute probabilities of actual interest to decision makers, *Int. J. Technol. Assess. Health Care* **17** (2001), pp. 17–26.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.*
***If stratified, skip to 2b4.9***

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**The following results are for the HWR measure.**

**Medicine Cohort Model Discrimination**

| Indices | 2007-2008 Development Sample | 2007-2008 Validation Sample | 2009 Validation Sample | 2015 HWR Data for Public Reporting |
|---|---|---|---|---|
| Number of hospital stays | 3,085,962 | 3,082,357 | 3,032,518 | 2,864,028 |
| Number of hospitals | 4,954 | 4,946 | 4,908 | 4,713 |
| Unadjusted readmission rate | 18.0% | 18.0% | 18.1% | 17.1% |
| Discrimination - Predictive Ability (lowest decile %, highest decile %) | 9 – 34 | 9 – 34 | 7 – 36 | 9-33 |
| Discrimination – c statistic | 0.640 | 0.641 | 0.663 | .643 |

**Surgical Cohort Model Discrimination**

| Indices | 2007-2008 Development Sample | 2007-2008 Validation Sample | 2009 Validation Sample | 2015 HWR Data for Public Reporting |
|---|---|---|---|---|
| Number of hospital stays | 2,208,753 | 2,208,482 | 2,109,292 | 1,695,227 |
| Number of hospitals | 4,354 | 4,353 | 4,232 | 4,031 |
| Unadjusted readmission rate | 12.6% | 12.6% | 12.6% | 11.1% |
| Discrimination - Predictive Ability (lowest decile %, highest decile %) | 4 – 27 | 4 – 27 | 3 – 30 | 5-27 |
| Discrimination – c statistic | 0.675 | 0.675 | 0.699 | 0.675 |

**Cardiorespiratory Cohort Model Discrimination**

| Indices | 2007-2008 Development Sample | 2007-2008 Validation Sample | 2009 Validation Sample | 2015 HWR Data for Public Reporting |
|---|---|---|---|---|
| Number of hospital stays | 1,396,562 | 1,396,855 | 1,331,539 | 1,144,451 |
| Number of hospitals | 4,810 | 4,806 | 4,718 | 4,596 |
| Unadjusted readmission rate | 21.1% | 21.2% | 21.4% | 19.5% |
| Discrimination - Predictive Ability (lowest decile %, highest decile %) | 11 – 37 | 11 – 37 | 9 – 40 | 10-35 |
| Discrimination – c statistic | 0.630 | 0.631 | 0.657 | 0.636 |

**Cardiovascular Cohort Model Discrimination**

| Indices | 2007-2008 Development Sample | 2007-2008 Validation Sample | 2009 Validation Sample | 2015 HWR Data for Public Reporting |
|---|---|---|---|---|
| Number of hospital stays | 860,485 | 861,925 | 809,520 | 707,529 |
| Number of hospitals | 4,702 | 4,703 | 4,641 | 4,438 |
| Unadjusted readmission rate | 15.2% | 15.2% | 15.4% | 14.4% |
| Discrimination - Predictive Ability (lowest decile %, highest decile %) | 5 – 31 | 6 – 30 | 5 – 33 | 7-31 |
| Discrimination – c statistic | 0.657 | 0.656 | 0.680 | 0.658 |

**Neurology Cohort Model Discrimination**

| Indices | 2007-2008 Development Sample | 2007-2008 Validation Sample | 2009 Validation Sample | 2015 HWR Data for Public Reporting |
|---|---|---|---|---|
| Number of hospital stays | 461,225 | 461,262 | 452,743 | 432,573 |
| Number of hospitals | 4,699 | 4,686 | 4,609 | 4,426 |
| Unadjusted readmission rate | 14.7% | 14.7% | 14.6% | 13.1% |
| Discrimination - Predictive Ability (lowest decile %, highest decile %) | 8 – 27 | 8 – 26 | 6 – 29 | 8-26 |
| Discrimination – c statistic | 0.614 | 0.613 | 0.646 | 0.622 |

**The following results are for the ACR measure.**

**Medicine Cohort Model Discrimination**

| Indices | 2014-2015 Development Sample | 2014-2015 Validation Sample |
|---|---|---|
| Number of hospital stays | 639,278 | 639,625 |
| Number of ACOs | 449 | 449 |
| Unadjusted readmission rate | 16.56% | 16.50% |
| Discrimination -Predictive Ability (lowest decile %, highest decile %) | 8 -- 33 | 8 -- 33 |

| | | |
|---|---|---|
| Discrimination – c statistic | 0.647 | 0.647 |

**Surgical Cohort Model Discrimination**

| Indices | 2014-2015 Development Sample | 2014-2015 Validation Sample |
|---|---|---|
| Number of hospital stays | 383,190 | 382,651 |
| Number of ACOs | 449 | 448 |
| Unadjusted readmission rate | 10.85% | 10.72% |
| Discrimination -Predictive Ability (lowest decile %, highest decile %) | 3 -- 26 | 3 -- 26 |
| Discrimination – c statistic | 0.694 | 0.694 |

**Cardiorespiratory Cohort Model Discrimination**

| Indices | 2014-2015 Development Sample | 2014-2015 Validation Sample |
|---|---|---|
| Number of hospital stays | 249,649 | 249,992 |
| Number of ACOs | 449 | 449 |
| Unadjusted readmission rate | 18.90% | 18.96% |
| Discrimination -Predictive Ability (lowest decile %, highest decile %) | 10 -- 34 | 10 -- 35 |
| Discrimination – c statistic | 0.638 | 0.639 |

**Cardiovascular Cohort Model Discrimination**

| Indices | 2014-2015 Development Sample | 2014-2015 Validation Sample |
|---|---|---|
| Number of hospital stays | 151,907 | 151,790 |
| Number of ACOs | 449 | 448 |
| Unadjusted readmission rate | 13.72% | 13.85% |
| | 7 -- 30 | 7 -- 30 |

| | | |
|---|---|---|
| Discrimination - Predictive Ability (lowest decile %, highest decile %) | | |
| Discrimination – c statistic | 0.66 | 0.66 |

**Neurology Cohort Model Discrimination**

| Indices | 2014-2015 Development Sample | 2014-2015 Validation Sample |
|---|---|---|
| Number of hospital stays | 92,759 | 92,503 |
| Number of ACOs | 448 | 448 |
| Unadjusted readmission rate | 12.97% | 12.95% |
| Discrimination - Predictive Ability (lowest decile %, highest decile %) | 7 -- 26 | 7 -- 26 |
| Discrimination – c statistic | 0.63 | 0.63 |

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**The following results are for the HWR measure.**

| Indices | 2007-2008 Development Sample | 2007-2008 Validation Sample | 2009 Sample |
|---|---|---|---|
| Medicine Cohort | (0, 1) | (0.011, 1.006) | (0.132, 1.118) |
| Surgical Cohort | (0, 1) | (-0.012, 0.995) | (0.104, 1.076) |
| Cardiorespiratory | (0, 1) | (0.010, 1.006) | (0.193, 1.184) |
| Cardiovascular Cohort | (0, 1) | (-0.019, 0.993) | (0.145, 1.109) |
| Neurology Cohort | (0, 1) | (-0.036, 0.982) | (0.201, 1.163) |

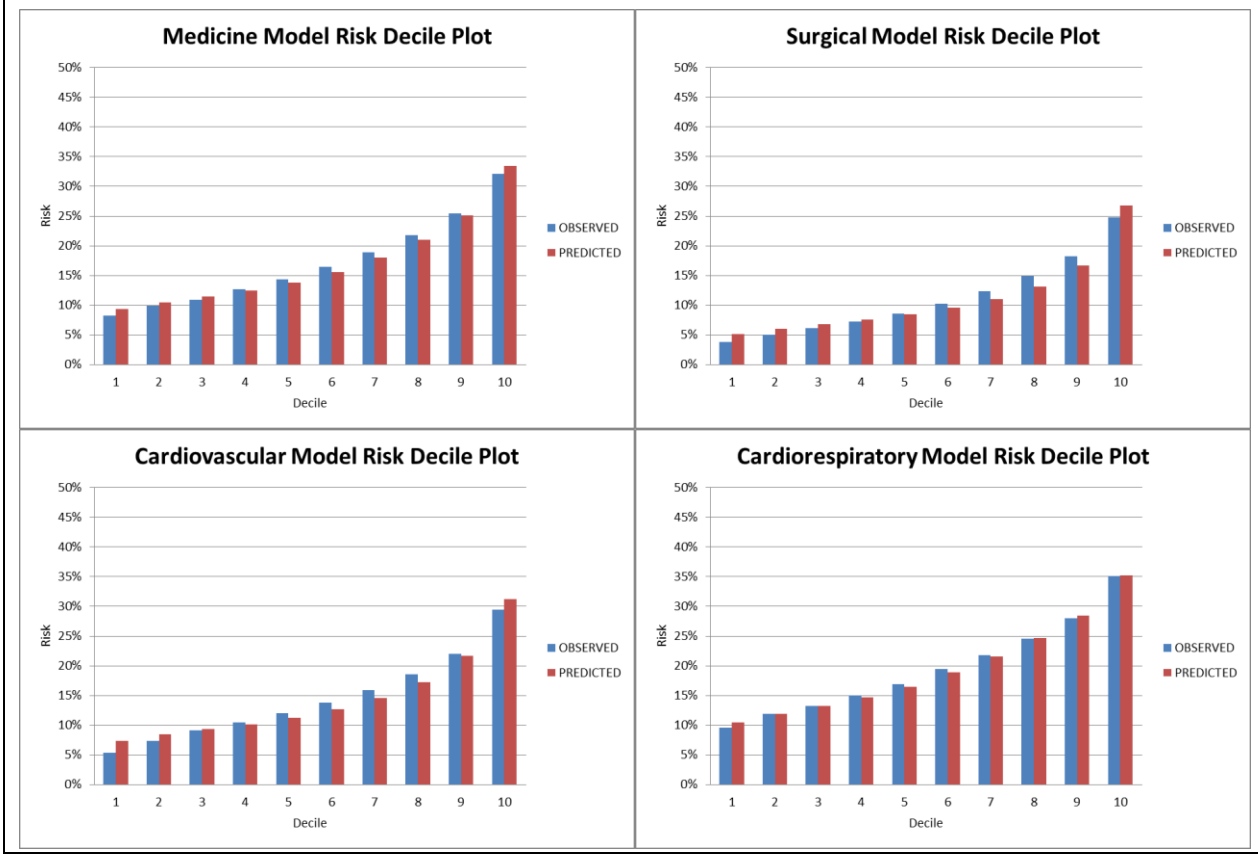
**The following results are for the ACR measure.**

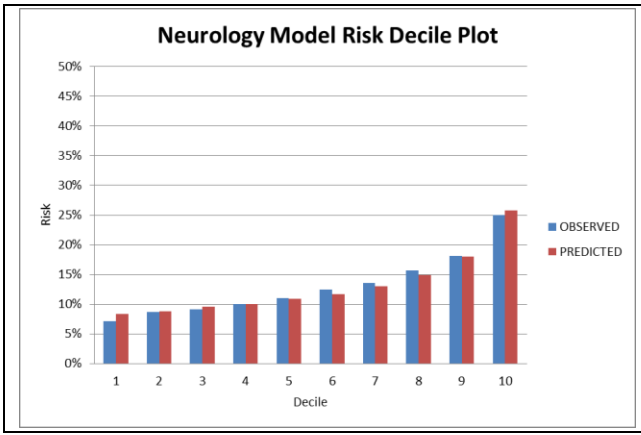
| Indices | 2014-2015 Development Sample | 2014-2015 Validation Sample |
|---|---|---|
| Medicine Cohort | (0, 1) | (-0.026, 0.987) |
| Surgical Cohort | (0, 1) | (-0.024, 0.995) |
| Cardiorespiratory Cohort | (0, 1) | (0.017, 1.008) |
| Cardiovascular Cohort | (0, 1) | (0.005, 0.995) |
| Neurology Cohort | (0, 1) | (-0.031, 0.993) |

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

**HWR**

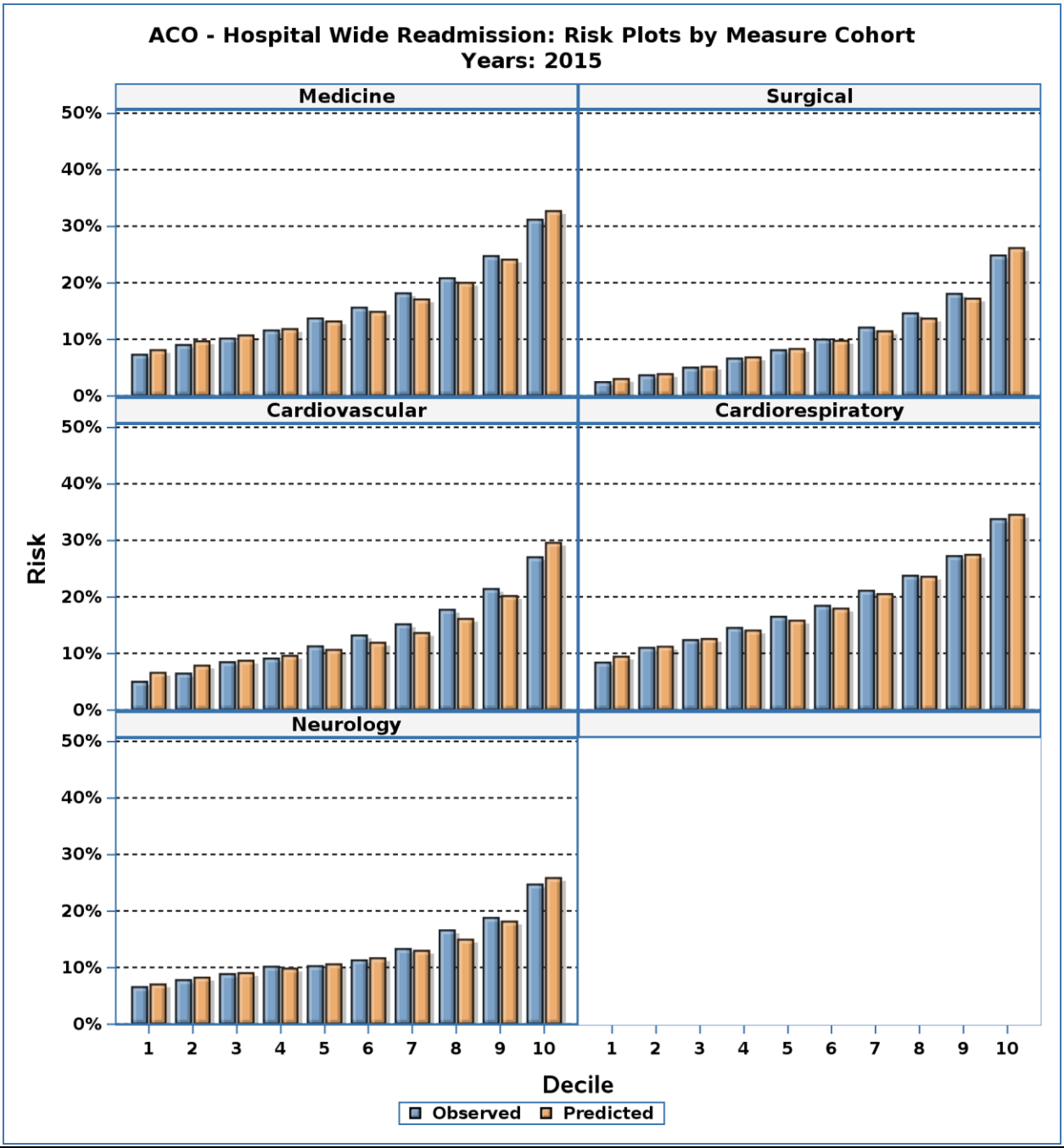
The risk decile plot is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot showing the distributions for Medicare FFS data from July 2013 to June 2014.




Neurology Model Risk Decile Plot
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Risk
1
2
3
4
5
6
7
8
9
10
Decile
OBSERVED
PREDICTED

## ACR

The risk decile plot is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot showing the distributions for Medicare ACOs in 2015. The following results are for the ACR measure.



**2b4.9. Results of Risk Stratification Analysis**:

| |
|---|
| N/A |

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i.*e., what do the results mean and what are the norms for the test conducted*)

**HWR**

*Discrimination Statistics*
The C-statistics indicate fair discrimination for each of the models in **Datasets 1 and 2**. Each of the models indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

*Calibration Statistics*
*Over-fitting (Calibration γ0, γ1)*
If the γ0 in the validation samples are substantially far from zero and the γ1 is substantially far from one, there is potential evidence of over-fitting. The calibration values close to 0 at one end and close to 1 to the other end indicate good calibration of each of the models.

*Risk Decile Plots*
Higher deciles of the predicted outcomes are associated with higher observed outcomes, which show a good calibration of the model. This plot indicates good discrimination of the model and good predictive ability.

*Overall Interpretation*
Interpreted together, our diagnostic results demonstrate that the risk-adjustment model adequately controls for differences in patient characteristics (case mix).

**ACR**

*Discrimination Statistics*
The C-statistics indicate fair discrimination for each of the models in **Development Sample and Validation Sample**. Each of the models indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

*Calibration Statistics*
The calibration values close to 0 at one end and close to 1 to the other end indicate good calibration of each of the models.

*Risk Decile Plots*
Higher deciles of the predicted outcomes are associated with higher observed outcomes, which show a good calibration of the model. This plot indicates good discrimination of the model and good predictive ability.

*Overall Interpretation*
Interpreted together, above diagnostic results demonstrate that the risk-adjustment model adequately controls for differences in patient characteristics (case mix).

**2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

**HWR**

For public reporting of the measure, CMS characterizes the uncertainty associated with the RSRR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSRR's interval estimate does not include the national observed readmission rate (is lower or higher than the rate), then CMS is confident that the hospital's RSRR is different from the national rate, and describes the hospital on the Hospital Compare website as "better than the U.S. national rate" or "worse than the U.S. national rate." If the interval includes the national rate, then CMS describes the hospital's RSRR as "no different than the U.S. national rate" or "the difference is uncertain." CMS does not classify performance for hospitals that have fewer than 25 cases in the one-year period.

**ACR**

In order to be eligible to share in any savings generated, an ACO must meet the established quality performance standard that corresponds to its performance year. In the first performance year of their first agreement period, ACO's satisfy the quality performance standard when they completely and accurately report on all quality measures (pay for reporting). The ACR measure is phased in to pay for performance in an ACO's third performance year (i.e. the ACR is pay for reporting for the first two performance years).

For measures that are pay-for-performance, quality scoring will be based on the ACO's level of performance on each measure compared to a benchmark. Centers for Medicare & Medicaid Services (CMS) establishes quality performance benchmarks prior to the reporting period for which they apply and are set for 2 years. CMS established 2016/2017 benchmarks using all available and applicable 2012, 2013 and 2014 Medicare fee-for-service (FFS) data. All of the

quality measure benchmarks were calculated using ACO, group practice and individual physician data aggregated to the TIN level and included if there were at least 20 cases. Quality data for ACO's, providers or group practices that did not satisfy the reporting requirements of the Shared Savings Program or PQRS were not included in calculation of the benchmarks.

A quality performance benchmark is the performance rate an ACO must achieve to earn the corresponding quality points for each measure. Below, we show the ACR benchmarks for each percentile, starting with the 30th percentile (corresponding to the minimum attainment level) and ending with the 90th percentile (corresponding to the maximum attainment level). For measures that are pay-for-performance, quality scoring will be based on the ACO's level of performance on each measure. After the second performance year of their first agreement period (i.e., the ACR is pay for reporting for the first two performance years), an ACO will earn quality points for the ACR measure on a sliding scale based on level of performance. Performance below the minimum attainment level (the 30th percentile) for a measure will receive zero points for that measure; performance at or above the 90th percentile of the quality performance benchmark earns the maximum points available for the measure. The sliding scale measure scoring approach is below.

2016/2017 Reporting Year ACO, Risk-Standardized, All-Condition Readmission Measure Benchmarks
30th perc. 15.32
40th perc. 15.19
50th perc. 15.07
60th perc. 14.97
70th perc. 14.87
80th perc. 14.74
90th perc. 14.54

Sliding Scale Measure Scoring Approach
ACO Performance Level Quality points
90+ percentile benchmark: 2.00 points
80+ percentile benchmark: 1.85 points
70+ percentile benchmark: 1.70 points
60+ percentile benchmark: 1.55 points
50+ percentile benchmark: 1.40 points
40+ percentile benchmark:1.25 points
30+ percentile benchmark: 1.10 point
<30 percentile benchmark: No points

To assess meaningful differences in performance, we present histograms of ACO performance in calendar years 2014 and 2015. ACOs may receive bonus points on the SSP care coordination domain, which demonstrated the second highest level of overall improvement. This achievement is significant given the difficulty of these measures. Up to four improvement points are available for each domain.
For more details about the benchmarking methodology, please review the Medicare Shared Savings Program Quality Measure Benchmarks for the 2016 and 2017 Reporting Years

https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/MSSP-QM-Benchmarks-2016.pdf.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

**HWR**

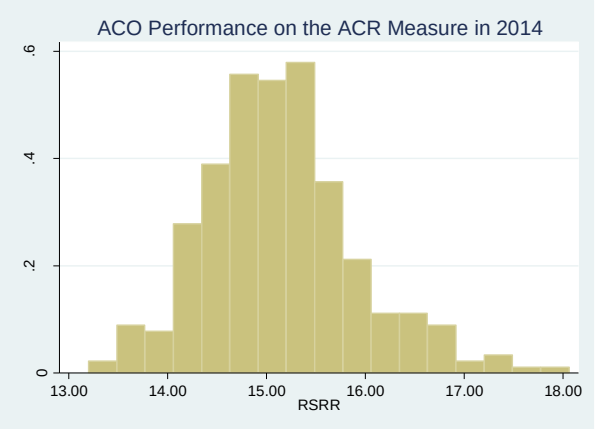
In the 2015 public reporting year (**Dataset 1**), out of 4,772 hospitals in the U.S., 178 performed "better than the U.S. national rate," 4,078 performed "no different from the U.S. national rate," and 337 performed "worse than the U.S. national rate." One hundred and seventy-nine hospitals were classified as "number of cases too small" (fewer than 25) to reliably tell how well the hospital is performing.

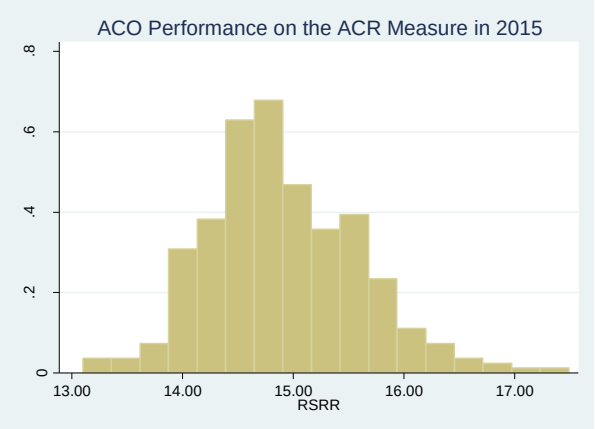
Note that this analysis included index admissions from July 2011 – June 2014 from the 2015 public reported data (**Dataset 1**). We used the planned readmission algorithm version 3.0 for measure calculation in these data. The planned readmission algorithm 4.0 will first be applied in the 2016 publically reported measure results.

**ACR**

The histograms below show ACO performance on the ACR measure in calendar years 2014 and 2015 using the 2014 and 2015 scoring datasets (**Dataset 5** and **Dataset 6**).

Distributions of ACR RSRRs




**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The variation in rates and number of performance outliers suggests that differences in the quality of care received across hospitals for the HWR measure remain, which support continued measurement in order to reduce variation.

Likewise, the variation in rates and number of performance outliers suggests that differences in the quality of care received across ACOs for the ACR measure remains, which supports continued measurement in order to reduce variation

## 2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

***If only one set of specifications, this section can be skipped.***

**Note**: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

N/A

**2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

## 2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

N/A

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

N/A

## 3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

**3a. Byproduct of Care Processes**

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

**3b. Electronic Sources**

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
**Attachment:**

**3c. Data Collection Strategy**

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
**IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.**
Administrative data are routinely collected as part of the billing process.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
There are no fees associated with the use of this measure.

## 4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

**4a. Accountability and Transparency**

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

**4.1. Current and Planned Use**

*NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.*

| Planned | Current Use (for current use provide URL) |
|---|---|
| | Public Reporting<br>Hospital Inpatient Quality Reporting (IQR) Program<br>http://cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html<br>Hospital Inpatient Quality Reporting (IQR) Program<br>http://cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html<br><br>Payment Program<br>Medicare SSP; Next Generation, and Pioneer ACO Model<br>https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/index.html |

**4a.1. For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

HWR

Program Name, Sponsor: Hospital Inpatient Quality Reporting (IQR) Program, Centers for Medicare and Medicaid Services (CMS)

Purpose: The Hospital Inpatient Quality Reporting (Hospital IQR) program was originally mandated by Section 501(b) of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. This section of the MMA authorized CMS to pay hospitals that successfully report designated quality measures a higher annual update to their payment rates. Initially, the MMA provided for a 0.4 percentage point reduction in the annual market basket (the measure of inflation in costs of goods and services used by hospitals in treating Medicare patients) update for hospitals that did not successfully report. The Deficit Reduction Act of 2005 increased that reduction to 2.0 percentage points.

In addition to giving hospitals a financial incentive to report the quality of their services, the hospital reporting program provides CMS with data to help consumers make more informed decisions about their health care. Some of the hospital quality of care information gathered through the program is available to consumers on the Hospital Compare website at: www.hospitalcompare.hhs.gov.

Geographic area and number and percentage of accountable entities and patients included:
The IQR program includes all IPPS non-federal acute care hospitals and Veteran Affairs (VA) hospitals in the United States. The number and percentage of accountable hospitals included in the program, as well as the number of patients included in the measure, varies by reporting year. For 2015 public reporting, the RSRR was reported for 4,772 hospitals across the U.S. The final index cohort includes 6,843,808 admissions.

ACR

The following programs and models use the ACO ACR measure, which has been in the Medicare SSP, Pioneer ACO Model, and the Next Generation ACO Model measure set since the Program's inception.

Medicare SSP: The Medicare SSP was established by Section 3022 of the Affordable Care Act. The SSP is a key component of the Medicare delivery system reform initiatives included in the Affordable Care Act and is a new approach to the delivery of health care. Through ACOs, the SSP facilitates coordination and cooperation among providers to improve the quality of care for Medicare FFS beneficiaries and lower the growth in Medicare expenditures. Eligible providers, hospitals, and suppliers may participate in the SSP by creating or participating in an ACO. The mean performance rate for SSP ACOs was 14.86 (range: 13.1-17.49) in 2015. As of January 2017, there are 480 SSP ACOs with over 9 million assigned beneficiaries across the 50 states, Puerto Rico, and Washington DC. An ACO may serve patients across multiple regions. ACOs include networks of individual practices, group practices, hospital/professional partnerships, hospitals employing ACO professionals, federally qualified health centers, rural health clinics, and critical access hospitals. An ACO may report multiple of these characteristics.

Pioneer ACO Model: The Pioneer ACO Model is designed for health care organizations and providers that are already experienced in coordinating care for patients across care settings. It will allow these provider groups to move more rapidly from a shared savings payment model to a population-based payment model on a track consistent with, but separate from, the Medicare SSP. It is designed to work in coordination with private payers by aligning provider incentives, which will improve quality and health outcomes for patients across the ACO, and achieve cost savings for Medicare, employers, and patients.
The mean performance rate for Pioneer ACOs was 15.41 (range: 13.98-16.71) in 2015. The Pioneer ACO Model began with 32 ACOs in 2012 and concluded December 31, 2016 with 8 ACOs participating. Pioneer ACOs are located across the US. https://innovation.cms.gov/initiatives/Pioneer-aco-model/

Next Generation ACO Model: The Next Generation ACO Model is an initiative for ACOs that are experienced in coordinating care for populations of patients. It will allow these provider groups to assume higher levels of financial risk and reward than are available under the current Pioneer Model and SSP. The goal of the Model is to test whether strong financial incentives for ACOs, coupled with tools to support better patient engagement and care management, can improve health outcomes and lower expenditures for Original Medicare fee-for-service (FFS) beneficiaries. https://innovation.cms.gov/initiatives/Next-Generation-ACO-Model/

Eighteen ACOs participated in the Next Generation ACO Model for the 2016 performance year, and twenty-eight ACOs are joining the Model for 2017.

ACOs voluntarily participate in the SSP/Pioneer ACO Model/Next Generation ACO Model following an application and CMS approval process. In these ACOs, the ACR measure reflects the RSRR at an ACO-level rather than a hospital level. ACOs vary substantially in composition and typically include multiple types of care delivery entities. For example, some ACOs are networks of group practices, some involve partnerships between ACO professionals and hospitals, and some are hospital-based ACOs. While ACOs may include hospitals as participants, ACOs are not required to have hospitals as participants. For instance, as of January 2017, only 38% of ACOs participating in the SSP involve partnerships between ACO professionals and hospitals.

In the SSP/Pioneer ACO Model/Next Generation ACO Model, the ACR measure is pay for reporting 2 years before phasing into pay for performance. ACOs receive full points for pay for reporting measures when they completely report data to CMS. For the ACR measure, no quality reporting is required by ACOs because CMS uses administrative claims to calculate the measure. The pay for reporting period of 2 performance years gives ACOs the opportunity to become familiar with the measure and understand their performance before CMS begins assessing ACO performance against a measure benchmark. Under pay for performance, ACO performance is compared to a benchmark that is calculated using all Medicare FFS data. In addition, benchmarks are set for 2 performance years to give ACOs a steady target for improving performance. ACOs may also receive quality improvement points in calculating the ACO Overall Quality Score, if they demonstrate a significant improvement in performance from one year to the next. The ACR measure performance is included in these quality improvement point calculations.

**4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A. This measure is currently publicly reported.

**4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for*

*implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A. This measure is currently publicly reported.

**4b. Improvement**

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

**4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)**
**Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:**

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

HWR
There has been significant progress in 30-day RSRR for unplanned, all-cause readmissions. The median 30-day RSRR decreased by 0.7 absolute percentage points from the 2013 public reporting period (median RSRR: 15.9%) to the 2015 public reporting period (median RSRR: 15.2%).

ACR
ACOs are already performing well compared to the national average. From 2012 to 2015, the ACO median RSRR decreased from 14.85 in 2012 to 14.83 in 2015. Please note, the number of ACOs participating in the SSP increased from 114 in 2012 to 397 ACOs in 2015. In 2015, 69.38% of Pioneer and SSP ACOs with two years of performance data improved their ACR rates. The maximum percent improvement for Pioneer and SSP ACOs was 12.3% with a mean of 3.4% from 2014 to 2015. Of those that improved, 33 SSP ACOs and 1 Pioneer Model ACO saw a significant enough improvement to receive additional points under the Quality Improvement Reward.

**4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

**4c. Unintended Consequences**

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

**4c.1. Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We did not identify any unexpected findings during the implementation of HWR measure. However, we are committed to monitoring this measure's use and assessing potential unintended consequences over time, such as the inappropriate shifting of care, increased patient morbidity and mortality, and other negative unintended consequences for patients.

Likewise, we did not identify any unexpected findings during the implementation of the ACR measure.

## 5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

**5. Relation to Other NQF-endorsed Measures**
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0329 : Risk-Adjusted 30-Day All-Cause Readmission Rate
0330 : Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following heart failure (HF) hospitalization
0505 : Hospital 30-day all-cause risk-standardized readmission rate (RSRR) following acute myocardial infarction (AMI) hospitalization.
0506 : Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following pneumonia hospitalization
0695 : Hospital 30-Day Risk-Standardized Readmission Rates following Percutaneous Coronary Intervention (PCI)
1551 : Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
1768 : Plan All-Cause Readmissions (PCR)
1891 : Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following chronic obstructive pulmonary disease (COPD) hospitalization

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

**5a. Harmonization**

The measure specifications are harmonized with related measures;
**OR**
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
**Are the measure specifications completely harmonized?**
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
This measure and the National Committee for Quality Assurance (NCQA) Plan All-Cause Readmissions (PCR) Measure #1768 are related measures, but are not competing because they don't have the same measure focus and same target population. In addition, both have been previously harmonized to the extent possible under the guidance of the National Quality Forum Steering Committee in 2011. Each of these measures has different specifications. NCQA's Measure #1768 counts the number of inpatient stays for patients aged 18 and older during a measurement year that were followed by an acute readmission for any diagnosis to any hospital within 30 days. It contrasts this count with a calculation of the predicted probability of an acute readmission. NCQA's measure is intended for quality monitoring and accountability at the health plan level. This measure estimates the risk-standardized rate of unplanned, all-cause readmissions to a hospital or ACO for any eligible condition within 30 days of hospital discharge for patients aged 18 and older. The measure will result in a single summary risk-adjusted readmission rate for conditions or procedures that fall under five specialties: surgery/gynecology, general medicine, cardiorespiratory, cardiovascular, and neurology. This measure is specified for evaluating hospital or ACO performance. However, despite these differences in cohort specifications, both measures under NQF guidance have been harmonized to the extent possible through modifications such as exclusion of planned readmissions. We did not include in our list of related measures any non-outcome (e.g., process) measures with the same target population as our measure. Because this is an outcome measure, clinical coherence of the cohort takes precedence over alignment with related non-outcome measures. Furthermore, non-outcome measures are limited due to broader patient exclusions. This is because they typically only include a specific subset of patients who are eligible for that measure (for example, patients who receive a specific medication or undergo a specific procedure).

**5b. Competing Measures**

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
**OR**
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
**Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)**

N/A

## Appendix

**A.1 Supplemental materials may be provided in an appendix.** All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** 2015_Measures_Reevaluation_Hospital-Wide_Readmission_AUS_Report_FINAL_508_Compliant_01-29-16_v1.0.pdf

## Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services (CMS)
Co.2 **Point of Contact:** Lein, Han, Lein.han@cms.hhs.gov, 410-786-0205-
Co.3 **Measure Developer if different from Measure Steward:** Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/CORE)
Co.4 **Point of Contact:** Karen, Dorsey, karen.dorsey@yale.edu, 203-764-5700-

## Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
**Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.**
The working group involved in the initial measure development is detailed in the original technical report available at www.qualitynet.org.

Our measure development team consisted of the following members:
Leora Horwitz, MD, MHS
Chohreh Partovian, MD, PhD
Zhenqiu Lin, PhD
Jeph Herrin, PhD
Jacqueline Grady, MS
Mitchell Conover, BA
Julia Montague, MPH
Chloe Dillaway, BA
Kathleen Bartczak, BA
Lisa Suter, MD, MHS
Joseph Ross, MD, MHS
Susannah Bernheim, MD, MHS
Harlan Krumholz, MD, SM
Elizabeth Drye, MD, SM

**Measure Developer/Steward Updates and Ongoing Maintenance**
Ad.2 **Year the measure was first released:** 2009
Ad.3 **Month and Year of most recent revision:** 01, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.6 **Copyright statement:** N/A
Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** N/A