NATIONAL QUALITY FORUM

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 #: 3612

Corresponding Measures:

De.2. Measure Title: Risk-Standardized Acute Cardiovascular-Related Hospital Admission Rates for Patients with Heart Failure under the Merit-based Incentive Payment System

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services (CMS)

De.3. Brief Description of Measure: Risk-standardized rate of acute, unplanned cardiovascular-related hospital admissions among Medicare Fee-for-Service (FFS) patients aged 65 years and older with heart failure (HF) or cardiomyopathy.

1b.1. Developer Rationale: Hospital admission rates are an effective marker of ambulatory care quality. Hospital admissions from the outpatient setting reflect a deterioration in patients' clinical status and as such reflect an outcome that is meaningful to both patients and providers. In addition, hospitalization increases potential exposure to iatrogenic injury and there are a number of increasingly recognized toxic effects of hospitalization (for example, sleep deprivation; poor nourishment; deconditioning from inactivity; confusion from medications; stress from mental exhaustion) leading to "post hospitalization syndrome [1]," which may contribute to the risk of readmission. Patients receiving optimal, coordinated high-quality care should use fewer inpatient services than patients receiving fragmented, low-quality care. Thus, high rates of hospitalization may, at least to some extent, signal poor quality of care or inefficiency in health system performance. There is evidence that outpatient clinicians can reduce HF patients' risk of hospitalizations in a variety of ways, including but not limited to accessible primary care, coordination across providers and across care settings, early attention to changes in clinical status, adoption of guideline-directed medical therapy, careful prescribing in patients with comorbidities, patient education, and support for self-management [2].

There is strong evidence that ambulatory care clinicians can influence admission rates by providing high quality of care [3-9]. For example, Brown et al. pointed to four ambulatory care focused Medicare Coordinated Care Demonstration programs that reduced hospitalizations for high-risk patients by 13-30 events per 100 beneficiaries per year (8-33% of hospitalizations). Brown et al. highlighted six program features that were associated with successfully reducing hospitalizations: 1) supplementing patient telephone calls with in-person meetings; 2) occasionally meeting in-person with providers; 3) acting as a communication hub for providers; 4) providing patients with evidence-based education; 5) providing strong medication management; and 6) providing comprehensive and timely transitional care after hospitalizations [3]. In addition, van Loenen et al. found that higher levels of provider continuity decreased the risk of avoidable hospitalizations for ambulatory care-sensitive conditions (ACSCs) and chronic diseases [8]. Hussey et al. [10] found that among Medicare

beneficiaries, greater continuity of care was associated with lower odds of hospitalization (OR=0.94, CI=0.93-0.95). Moreover, several studies have demonstrated positive impact of early follow-up after hospitalization to reduce readmissions for HF [11-14].

Thus, the anticipated net benefits of this unplanned hospital admission measure include, but are not limited to:

- Improved patient experience through harm prevention and reduction.
- Better education about HF management for patients and caregivers.
- Improved support for self-management of HF and efforts to build capacity to carry out treatment plans.
- Reduced emergency visits, observation stays, and hospital admissions for events caused by HF.
- Reduced rates of poor outcomes associated with HF (falls, pneumonia, mortality, cardiovascular events).
- Potential cost savings to Medicare, patients, and taxpayers.

Overall, this measure will provide the Centers for Medicare & Medicaid Services (CMS) with a valuable tool for assessing the performance of outpatient clinicians and groups of clinicians in the MIPS program.

References

1. Krumholz HM. Post-Hospital Syndrome — An Acquired, Transient Condition of Generalized Risk. New England Journal of Medicine. 2013;368(2):100-102.

2. Jackevicius CA, de Leon NK, Lu L, Chang DS, Warner AL, Mody FV. Impact of a Multidisciplinary Heart Failure Post-Hospitalization Program on Heart Failure Readmission Rates. The Annals of pharmacotherapy. 2015;49(11):1189-1196.

3. Brown RS, Peikes D, Peterson G, Schore J, Razafindrakoto CM. Six Features of Medicare Coordinated Care Demonstration Programs That Cut Hospital Admissions of High-Risk Patients. Health Affairs. 2012;31(6):1156-1166.

4. Dorr DA, Wilcox AB, Brunker CP, Burdon RE, Donnelly SM. The Effect of Technology-Supported, Multidisease Care Management on the Mortality and Hospitalization of Seniors. Journal of the American Geriatrics Society. 2008;56(12):2195-2202.

5. Levine S, Steinman BA, Attaway K, Jung T, Enguidanos S. Home care program for patients at high risk of hospitalization. The American journal of managed care. 2012;18(8):e269-e276.

6. Littleford A, Kralik D. Making a difference through integrated community care for older people. Journal of Nursing and Healthcare of Chronic Illness. 2010;2(3):178-186.

7. Sommers LS, Marton KI, Barbaccia JC, Randolph J. Physician, Nurse, and Social Worker Collaboration in Primary Care for Chronically Ill Seniors. Archives of Internal Medicine. 2000;160(12):1825-1833.

8. Van Loenen T, Faber MJ, Westert GP, Van den Berg MJ. The impact of primary care organization on avoidable hospital admissions for diabetes in 23 countries. Scandinavian journal of primary health care. 2016;34(1):5-12.

9. Zhang NJ, Wan TTH, Rossiter LF, Murawski MM, Patel UB. Evaluation of chronic disease management on outcomes and cost of care for Medicaid beneficiaries. Health Policy. 2008;86(2):345-354.

10. Hussey PS, Schneider EC, Rudin RS, Fox DS, Lai J, Pollack CE. Continuity and the Costs of Care for Chronic Disease Care Continuity and Costs for Chronic Disease Care Continuity and Costs for Chronic Disease. JAMA Internal Medicine. 2014;174(5):742-748.

11. Donaho EK, Hall AC, Gass JA, et al. Protocol-Driven Allied Health Post-Discharge Transition Clinic to Reduce Hospital Readmissions in Heart Failure. Journal of the American Heart Association. 2015;4(12):e002296.

12. Lee KK, Yang J, Hernandez AF, Steimle AE, Go AS. Post-discharge Follow-up Characteristics Associated With 30-Day Readmission After Heart Failure Hospitalization. Medical Care. 2016;54(4):365-372.

13. Murtaugh CM, Deb P, Zhu C, et al. Reducing Readmissions among Heart Failure Patients Discharged to Home Health Care: Effectiveness of Early and Intensive Nursing Services and Early Physician Follow-Up. Health Services Research. 2017;52(4):1445-1472.

14. Ryan J, Kang S, Dolacky S, Ingrassia J, Ganeshan R. Change in Readmissions and Follow-up Visits as Part of a Heart Failure Readmission Quality Improvement Initiative. The American Journal of Medicine. 2013;126(11):989-994.e981.

S.4. Numerator Statement: The outcome for this measure is the number of acute cardiovascular-related admissions per 100 person-years at risk for admission during the measurement year.

S.6. Denominator Statement: This measure assesses the care provided to patients with heart failure by primary care providers and cardiologists.

Patients included in the measure (target patient population)

The target patient population for the outcome includes Medicare FFS patients aged 65 years and older with heart failure or cardiomyopathy.

Provider types included for measurement

- Primary care providers (PCPs): CMS designates PCPs as physicians who practice internal medicine, family medicine, general medicine, or geriatric medicine, and non-physician providers, including nurse practitioners, certified clinical nurse specialists, and physician assistants.
- Cardiologists: Cardiologists are covered by the measure because they provide overall coordination of care for patients with HF and manage the conditions that put HF patients at risk for admission due to acute cardiovascular-related conditions.

Outcome attribution

The measure begins by assigning each patient to the clinician most responsible for the patient's care, based on the pattern of outpatient visits with PCPs and relevant specialists. The patient can be assigned to a PCP, a cardiologist, or can be left unassigned. Patients who have had no Evaluation and Management (E&M) visits with a MIPS eligible clinician are excluded.

Step 1: A patient who is eligible for attribution is assigned to a cardiologist only if the cardiologist has been identified as "dominant." A cardiologist is considered "dominant" if they have two or more visits with the patient, regardless of how many visits that patient has with a PCP.

- There are two scenarios where a patient can be assigned to a PCP. First, if the patient has seen the PCP at least once but has no visits with a cardiologist, the patient is assigned to the PCP. The patient will then be assigned to the PCP with the highest number of visits as long as there are no relevant specialists who are considered "dominant." Second, if the patient has seen the PCP more than two or more times and has only one visit with a cardiologist, the patient is assigned to the PCP.
- If the patient has one visit each with a cardiologist and a PCP, the patient is assigned to the cardiologist.
- If the patient has one visit with a cardiologist and no visit with a PCP, the patient is assigned to the cardiologist.
- Finally, the patient will be unassigned if they only saw non-relevant specialists, if the patient has not seen a PCP and no "dominant" specialist can be identified, or if the patient has not had more than one visit with any individual PCP.

Step 2: Patients are then assigned at the Taxpayer Identification Number (TIN) level, which includes solo clinicians and groups of clinicians who have chosen to report their quality under a common TIN.

At the TIN level, patients are first assigned to the clinician (NPI/TIN) most responsible for their care (using the algorithm for individual clinician-level attribution above). Then, patients "follow" their attributed clinician to

the TIN of that clinician. Patients unassigned at the individual clinician level continue to be unassigned at the TIN level.

S.8. Denominator Exclusions: The measure excludes:

1. Patients without continuous enrollment in Medicare Part A and B for the duration of the measurement period.

2. Patients in hospice during the year prior to the measurement year or in hospice at the start of the measurement year.

3. Patients who have had a heart transplant, been on home inotropic therapy, or who have had a left ventricular assist device (LVAD) placed.

4. Patients with end stage renal disease (ESRD), defined as chronic kidney disease stage 5 or on dialysis.

5. Patients who had no E&M visits with MIPS eligible clinician.

De.1. Measure Type: Outcome

S.17. Data Source: Claims, Other

S.20. Level of Analysis: Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

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? Not applicable (N/A); this measure is not formally paired with another measure.

Preliminary Analysis: New Measure

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

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary

- The developer outlines a logic model depicting rates of admissions for patients with heart failure (HF) can be decreased through care coordination and continuity of care from outpatient providers.
- The developer cites evidence suggesting that outpatient clinicians can improve HF patients' risk of hospitalizations in a variety of ways, including but not limited to accessible primary care, coordination

across providers and across care settings, early attention to changes in clinical status, adoption of guideline-directed medical therapy, careful prescribing in patients with comorbidities, patient education, and support for self-management.

- One study found reductions in hospitalizations for high-risk patients by 13-30 events per 100 beneficiaries per year (8-33% of hospitalizations) due to various interventions, including supplementing patient telephone calls with in-person meetings; occasionally meeting in-person with providers; providing patients with evidence-based education; providing strong medication management; and providing comprehensive and timely transitional care after hospitalizations.
- Another study found that among Medicare beneficiaries, greater continuity of care was associated with lower hospitalization odds (OR=0.94, CI=0.93-0.95).

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm

BOX 1 – Yes à BOX 2 – Yes à PASS

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

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

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

- The developer reports data from Q4 2017 – Q3 2018 Medicare claims data for 1,846,193 Medicare FFS beneficiaries with HF.
- Across all tax identification numbers (TINs), which includes solo clinicians and groups of clinicians who have chosen to report their quality under a common TIN, the risk-standardized acute cardiovascular-related admission rate (RSCAR) measure scores ranged from 9.6 to 62.4 per 100 person-years, with a median of 24.8 and an interquartile range of 24.0 to 25.9. The mean RSCAR and standard deviation were 25.1 ± 2.4 admissions per 100 person-years.
- The developer reports in the Meaningful Differences section, that across the 10,760 TINs with at least 21 HF patients, RSCAR measure scores, including adjustment for the Agency for Healthcare Research and Quality Socieoeconomic status Index, ranged from 9.6 to 62.4 per 100 person-years, with a median of 24.9 and an IQR of 22.7 to 27.8. This indicates that after adjustment, half of the TINs had outcomes between 23 and 28 acute hospital admissions per 100 person years.
    - The 10th and 90th percentiles, representing the best and worst performers, had an admission rate of 20.9 and 30.9 respectively, representing deviations from the median: TINs in the 10th percentile (better performers) had 16% fewer admissions per 100-person years compared with the median, and TINs in the 90th percentile had 24% more admissions per 100-person years compared with the median.

Disparities

- The developer reports distributions of risk-standardized cardiovascular acute hospital admission rates by deciles and notes that these were generally similar across quartiles of the proportion of Medicare-Medicaid dual-eligible beneficiaries across TINs.
- Distribution of risk-standardized cardiovascular acute hospital admission rates by deciles, all TINs
    - Decile / Q1 of % dual (0.0 - 0.0) / Q2 of % dual (0.5 - 12.5) / Q3 of % dual (12.5 - 38.4) / Q4 of % dual (38.5 - 100.0)
    - 1 / 17.3 - 23.7 / 9.6 - 21.3 / 13.3 - 22.5 / 13.8 - 23.2

- 2 / 23.7 - 24.2 / 21.3 - 22.5 / 22.5 - 23.4 / 23.2 - 24.0
- 3 / 24.2 - 24.5 / 22.5 - 23.2 / 23.4 - 23.9 / 24.0 - 24.4
- 4 / 24.5 - 24.6 / 23.2 - 23.9 / 23.9 - 24.3 / 24.4 - 24.6
- 5 / 24.6 - 24.8 / 23.9 - 24.6 / 24.3 - 24.8 / 24.6 - 24.8
- 6 / 24.8 - 24.8 / 24.6 - 25.4 / 24.8 - 25.5 / 24.8 - 24.9
- 7 / 24.8 - 24.9 / 25.4 - 26.4 / 25.5 - 26.1 / 24.9 - 25.5
- 8 / 24.9 - 25.5 / 26.4 - 27.7 / 26.1 - 27.1 / 25.5 - 26.3
- 9 / 25.5 - 26.3 / 27.7 - 29.8 / 27.1 - 28.9 / 26.3 - 27.3
- 10 / 26.3 - 36.0 / 29.8 - 55.5 / 28.9 - 49.6 / 27.3 - 62.4

- Distribution of risk-standardized cardiovascular acute hospital admission rates by deciles, TINs with >= 32 patients (which was determined to yield a min. reliability estimate=0.5)
  - Decile / Q1 of % dual (0.0 - 0.0) / Q1 of % dual (0.0 - 7.7) / Q2 of % dual (7.7 - 15.1) / Q3 of % dual (15.2 - 28.9) / Q4 of % dual (28.9 - 100.0)
  - 1 / 17.3 - 23.7 / 9.6 - 20.1 / 15.1 - 20.8 / 14.5 - 21.0 / 13.3 - 20.4
  - 2 / 23.7 - 24.2 / 20.1 - 21.4 / 20.8 - 22.4 / 21.0 - 22.3 / 20.4 - 22.0
  - 3 / 24.2 - 24.5 / 21.4 - 22.4 / 22.4 - 23.6 / 22.3 - 23.4 / 22.0 - 23.1
  - 4 / 24.5 - 24.6 / 22.4 - 23.2 / 23.6 - 24.8 / 23.4 - 24.4 / 23.1 - 24.0
  - 5 / 24.6 - 24.8 / 23.2 - 24.3 / 24.8 - 25.9 / 24.4 - 25.4 / 24.0 - 25.0
  - 6 / 24.8 - 24.8 / 24.3 - 25.4 / 25.9 - 27.1 / 25.5 - 26.6 / 25.0 - 26.1
  - 7 / 24.8 - 24.9 / 25.4 - 26.6 / 27.1 - 28.3 / 26.6 - 28.1 / 26.1 - 27.4
  - 8 / 24.9 - 25.5 / 26.6 - 28.3 / 28.3 - 29.8 / 28.1 - 29.5 / 27.5 - 29.0
  - 9 / 25.5 - 26.3 / 28.3 - 30.5 / 29.9 - 32.5 / 29.6 - 32.2 / 29.0 - 31.5
  - 10 / 26.3 - 36.0 / 30.5 - 50.7 / 32.5 - 55.5 / 32.2 - 49.6 / 31.5 - 62.4

Questions for the Committee:

- *The Standing Committee should consider whether there a gap in care that warrants a national performance measure?*

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

Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

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."

- Data shows substantial variance but reported data apparently include all clinicians/TINS. Would like to see variance for TINS with >21 and >30 patients.
- Responding to the question posed by NQF staff, incorporating data from their logic model, if feasible, could change results.
- Developer referenced literature review of articles that showed strong evidence that ambulatory care clinicians can reduce hospitalizations by continuity of care by providers and various interventions such as strong medication management, in-person visits, and evidence-based patient education.
- No concerns
- Did not see where PRO was included in measurement

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?

- Data shows substantial variance but reported data apparently include all clinicians/TINS. Would like to see variance for TINS with >21 and >30 patients.
- Most clinicians would agree there is a gap in care. The question is why it is not demonstrated.
- Yes. Medicare FFS claims of HF beneficiaries from Q4 2017-Q3 2018 showed the risk-standardized acute cardiovascular-related admission rate (RSCAR) measure scores varied across all tax identification numbers. The RSCAR measure score ranged from 9.6 to 62.4 per 100 person years
- No concerns
- Disparities not specifically addressed; performance was based on either group or individual provider. Disparities known to some populations would impact the measurement, especially for an individual provider who serves a vulnerable population.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

2c. For composite measures: empirical analysis support composite approach

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: NQF Scientific Methods Panel
Reliability: H-0; M-5; L-3; I-0 (Pass)
Validity: H-0; M-6; L-2; I-0 (Pass)

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

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

Reliability

- The developer conducted reliability testing at the performance score level:
    - The developer performed a signal-to-noise analysis.
    - The minimum HF patient sample size for TINs needed to achieve minimum reliability scores of 0.4 and 0.5 was determined to be between 21 and 32.
    - The developer noted that a minimum reliability of 0.4 was achieved for TINs with at least 21 HF patients. At this threshold, reliability scores for TINs ranged from 0.40 to nearly 1.0, with a median value of 0.600 (IQR 0.481-0.778). With the 21-patient volume minimum, the measure included 23.9% of clinician groups; however, 88.9% of the patients, 91.3% of the admissions, and 69.8% of clinicians, who reported under these TINs.
- The SMP members agreed that approach is appropriate, but they raised several concerns including:

    - The reliability tests are not conducted and presented for clinical groups and individual clinicians separately.
    - Unit of analysis is not clear throughout the result section.
    - Reliabilities are much higher (median = 0.60) among providers with at least 21 eligible cases. However, only 24% of providers had this many cases which again sounds a bit low.
- The developer provided responses to the SMP concerns, noting that under the Merit-based Incentive Payment System (MIPS), clinicians annually select whether to report as individuals, as part of a group, or as both. The group includes both solo clinicians (i.e., clinicians opting not to report with other clinicians under MIPS) and groups of clinicians who have chosen to report their quality under a common tax identification number (TIN). Therefore, testing results include both individual clinicians and clinician groups, consistent with how the MIPS program evaluates quality.
- Among TINs with a case volume of at least 21 HF patients (when reliability of 0.4 is reached), 31.8 percent were solo clinicians. Further, the 21 minimum case volume was established to reach the reliability threshold of 0.4, and that the MIPS program will set the minimum case volume during rulemaking.
- The SMP acknowledged the developer's response and passed the measure on reliability.

Validity

- The developer conducted face validity of the measure score, which is the minimum acceptable testing for new measures.
- Face validity was demonstrated through assessment from external groups (a technical expert panel [TEP] and clinician committee) and from the use of established measure development guidelines.
    - Of 17 TEP members who were active through the end of the project, 12 responded. The majority of the respondents, 10/12 or 83%, moderately or somewhat agreed that the MIPS HF measure can be used to distinguish good from poor quality of care.
    - Of the 13 Clinician Committee members who responded to the survey, 11/13 or 85%, strongly, moderately, or somewhat agreed that the MIPS HF measure can be used to distinguish good from poor quality of care.
- For the risk adjustment model, the developer adjusted for 30 risk variables, including AHRQ SES Index. The R-squared for the model with demographic and clinical risk factors was 0.073 in the Development HF Full Sample and 0.072 in the Validation HF Full Sample, indicating that the model explains 7.3% and 7.2% of the variation, respectively, in admission rates. The Q4 2017 – Q3 2018 Medicare HF Full Sample R-squared after adding the AHRQ SES Index to the model was unchanged (0.073).
- The SMP members raised some concerns about the clarity of measure specifications, including attribution, exclusions (e.g., patients in hospice, patients with no E&M visits, CKD-4), and whether HF is the primary diagnosis.
- The SMP members in general thought the face validity is established adequately; however, there was some concern expressed related to the potential response bias, since not all technical expert panel (TEP) members responded to the survey for face validity.
- The SMP members thought the risk adjustment model is adequate, although they noticed that indicators of heart failure severity are not included, and the model did not appear to account for the repeated measures impact (i.e., single patient with multiple admissions vs. multiple patients with single admission). There are also questions about why race and dual-eligible are not included as they both can affect the outcome.

- The developer provided responses to the SMP concerns, noting that this was a multi-year effort and not all group members were active throughout the time of development. Only those remained responded to the survey. With respect to the feedback, this measure underwent multiple revisions with input from both groups (TEP and clinician committee), such as excluding some high-risk heart failure patients due to clustering of patients to certain clinicians, conditions, or devices (e.g., pacemakers, end stage renal disease, systolic heart failure), which all could lead to higher readmissions.
- The developer also stated that due to CMS request, the measure is not risk adjusted for race.
- The SMP acknowledged the developer's response passed the measure on validity.

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, 2b, and 2c)

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?

- Attribution based on 1 visit, 2 for cardiologists, might be questioned.
- No concerns.
- Signal-to-noise reliability analysis done. Median score 0.60 for volume of at least 21 patients.
- No concerns
- case mix needs to include race/ethnicity

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

- Reliability adequate at >30 patient level. Marginal at >21 level.
- It concerns me that there are no high ratings.
- No but agree with SMP recommendation to include only patients with primary discharge diagnosis of heart failure
- No concerns
- None

2b1. Validity -Testing: Do you have any concerns with the testing results?

- New measure. Face Validity only. Adequate
- It concerns me that none of the SMP evaluations are high.
- No. Face validity included Technical Expert Panel and Clinician Committee. Both groups had majority agreement with 2 questions
- No concerns
- None

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?

- Concerned about specifications. Included admission diagnoses include mechanical breakdowns of implanted devices (e.g. line 445-495 in data dictionary Tab 4). Would like discussion of whether PCP or cardiologist should be held responsible for these admissions.
- I think this measure and others like it are prone to bias. I am not sure not adjusting for race takes care of the problem.
- Yes. SRFs variables tested but only AHRQ SES included
- No concerns
- Hospice, heart transplant and renal failure patients excluded; seems reasonable

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.

- ALL data elements are in defined fields in a combination of electronic sources
- This measure uses administrative claims data and, as such, imposes no data collection burden to measure entities.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

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?

- claims based measure. no issues.
- No concerns as written.
- No concerns. No data collection burden to hospitals are providers since electronic sources using administrative claims and enrollment data
- No concerns
- race, ethnicity, and other SDoH that could impact measurement

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** ☐ **UNCLEAR**

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

Accountability program details

- The developer reports that this measure is not currently publicly reported or used in an accountability application. However, CMS may propose this measure for use under the Merit-based Incentive Payment System.
- The developer states that the intended audience are primary care and cardiology ambulatory care practices. The timeline for implementation has not been finalized at this time.

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

- To garner feedback on the development of the measure, the developer convened a national TEP, which included representatives of the measured entities and patients covered by the measure to ensure the measure is as meaningful as possible to all stakeholders. The developer provided performance results and data to TEP members periodically for their review and input.
- The developer further convened a Clinician Committee of professional society representatives and front-line clinicians from rural and/or underserved communities. The Clinician Committee provided more detailed input during the measure development process.
- During measure development, the developer reports that feedback was obtained with respect to cohort definition (e.g., exclusion of patients with heart transplant or on home inotropic therapy, exclusion of patients with end stage renal disease), attribution algorithm (e.g., single versus multiple providers), outcome definition (e.g., 10-day buffer period after admission), and risk adjustment (e.g., adjustment for AHRQ SES Index but not for dual eligibility).
- In response to this feedback, the developer revised the measure with respect to cohort definition, outcome, and risk adjustment. Some of the changes specifically made included exclusion of patients with end stage renal disease and exclusion of patients on home inotropic therapy from the cohort, and adjustment for systolic heart failure.

Additional Feedback:

- This measure was reviewed by the Measure Applications Partnership (MAP) for the 2020-2021 cycle.
- MAP did not recommend the measure for rulemaking with potential for mitigation. Mitigation points were: 1) NQF endorsement and 2) an analysis of the appropriateness of the risk adjustment for clinicians with higher caseloads of patients with more complicated or severe heart failure.
- The MAP noted that while the measure raises concerns that the risk adjustment may not adequately account for advanced heart failure stages, the measure also centers on an important need. As the MAP discussed, these points will be addressed by the NQF endorsement process.
- Based on the MAP feedback, the measure is being submitted for NQF endorsement. The developer states that the measure accounts for case-mix and heart failure severity in several ways: 1) excludes patients at advanced stages of heart failure, such as those with implanted left ventricular assist device (LVAD), those who receive home inotropic therapy, or those with prior heart transplant or with end stage renal disease; 2) risk adjusts for AICDs (defibrillators); 3) risk adjusts for systolic heart failure (which portends a poor prognosis); 4) risk adjusts for comorbidities including chronic kidney disease, and for frailty/disability.
- Additionally, the developer will continue to evaluate the risk model during regular measure maintenance; notably, the model performs well as currently specified.

***Questions for the Committee**:*

- *Does the Standing Committee have any concerns related to the potential use of the measure?*

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

- This measure is not currently in use. However, the developer states that "the primary goal of the measure is to provide information necessary to implement focused quality improvement efforts. Providers could use the measure information to implement practice improvements, such as those outlined in the Evidence attachment.

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

- This is a new measure and not currently in use. The developer does not report any unexpected findings.

Potential harms

- The developer does not report any unexpected findings.

Additional Feedback:

- The developer does not report any unexpected findings.

Questions for the Committee*:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

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?

- Feedback mechanism described in documentation. Usable by clinician, perhaps (would like some direct commentary from hospitals on how they use the reports). Limited predicted power of risk adjuster suggests either high variability in practices affecting outcomes or inherent variability in admission. Potential patient use of measure is minimal.
- New measure.
- New measure
- No concerns
- reported by CMS; no concerns

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.

- No obvious harms.
- The measure has the potential to provide actionable data. However, it seems likely that providers with sicker patients will have difficulty scoring well.
- New measure.
- No concerns
- Again, need to include race, ethnicity and SDoH for developing improvement processes.

Criterion 5: Related and Competing Measures

Related measures

- 2886 : Risk-Standardized Acute Admission Rates for Patients with Heart Failure

Harmonization

- The developer notes that this MIPS HF admission measure is adapted from the ACO HF admission measure, which was implemented in the Medicare Shared Savings Program in 2015.
- The developer states that there are three main ways that this measure differs from the ACO measure. The developer also provides supporting rationale for each of the differences below:
    - Cohort - Added cardiomyopathy as a cohort-qualifying condition
    - Outcome - Narrowed the outcome to focus on admissions whose risk can be reduced by clinicians/groups providing high-quality ambulatory care, so that the measure can be used to assess ambulatory (rather than ACO-wide) care quality.
    - Risk-adjustment – Added a social risk factor to the risk-adjustment model – namely, the AHRQ SES Index

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- Several related measures for different entities. Should be discussed.
- No.
- This measure is adapted from ACO HF admission measure but noted differences in the cohort, outcome focus and this measure added SRF adjustment
- No concerns
- none known

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 6/10/2021

The Federation of American Hospitals

The Federation of American Hospitals (FAH) appreciates the opportunity to comment on this measure. FAH agrees that measuring the frequency of admissions for patients with heart failure enables clinicians to understand where quality improvement efforts may be needed but does not support this measure for accountability uses due to several factors, including: there is insufficient evidence to support attribution to clinician groups; the minimum sample size and reliability threshold remain too low; and additional risk factors in the risk adjustment model are needed.

The FAH does not believe that it is appropriate to attribute these admissions to clinician groups. We were unable to find any data and empirical evidence to demonstrate that groups can meaningfully influence unplanned admissions for patients with heart failure. A practice's improvement in avoiding unplanned admissions must be based on its ability to leverage one or more structures or processes of care.

The FAH is concerned that while the median reliability score was 0.60 for practices with at least 21 patients, the range was from 0.401 to 0.995. The FAH believes that the developer must increase the minimum sample size to a higher number to produce a minimum reliability threshold of sufficient magnitude (e.g. 0.7 or higher). Ensuring that the resulting performance scores produce information that would not misrepresent the quality of care provided by a group is imperative and while an increase in the sample size would result in a decrease in the number of groups to which the measure would apply, we believe that it would still be a considerable number of patients with heart failure that would continue to be factored into the measure.

The FAH applauds the developer for including social risk factors within the risk adjustment model and strongly advocates that dual eligibility also be included since it was a strong predictor of whether a patient would be admitted. If the desire is to develop measures that can be used in other programs that may not include an adjustment for complex patients, then it becomes imperative that all variables that are determined to be predictors that are outside of the control of a group be included.

American Medical Association

The American Medical Association (AMA) appreciates the opportunity to comment on this measure. We strongly believe that while it is useful to understand the rate of admissions for patients with heart failure particularly for quality improvement, measures used in accountability programs must be based on strong evidence, actionable to ensure that improvements can be driven by those held accountable, and proven to be reliable and valid at the levels to which the measure is attributed.

The AMA is concerned with the lack of evidence to support attribution of the measure at the individual physician level. Attribution must be determined based on evidence that the accountable unit is able to meaningfully influence the outcome, which aligns with the National Quality Forum (NQF) report, Improving Attribution Models. We believe that there are several concerns that are not adequately addressed including:

- Heart failure patients are often cared for by more than one cardiologist.

- More clarity around the definition of inpatient vs. outpatient providers (e.g., cardiologists) would be helpful.

- Many practices in large organizations comprise both primary and specialty practices and therefore it is not entirely clear how attribution might be determined.

- This may be of concern, for example, with Advanced Practice Practitioners who are often considered primary care, but may also be in a cardiology practice. In this scenario, if a cardiology-specific APP has the most patient touchpoints, attribution could fall within primary care while in fact the cardiology practice is driving costs.

- Another example is an electrophysiologist who sees an appropriately referred patient for a device — and sees that patient twice in one year (e.g., the initial consultation, a follow-up visit) — she will now “own” the HF care for the year over the primary care provider, based on attribution logic.

We are also disappointed to see the minimum measure score reliability results of 0.401 using a minimum case number of 21 patients. We believe that measures must meet minimum acceptable thresholds of 0.7 for reliability.

The AMA supports and is encouraged to see that social risk factors were tested and will be included in the risk adjustment approach. We strongly recommend that dual eligibility be included in the adjustment since the results demonstrate that it is strongly predictive of an admission. We remain concerned that CMS continues to test social risk factors after assessment of clinical and demographic risk factors and it is unclear why this multi-

step approach is preferable. On review of the Evaluation of the NQF Trial period for Risk Adjustment for Social Risk Factors report, it is clear that the approaches to testing these data should be revised to strategies such as multi-level models or testing of social factors prior to clinical factors and that as access to new data becomes available, it may elucidate more differences that are unrelated to factors within a hospital's or physician's control. Additional testing that evaluates clinical and social risk factors at the same time or social prior to clinical variables rather than the current approach with clinical factors prioritized should be completed. This additional testing may provide support for inclusion of additional variables such as PCP density and further emphasize the need to include dual eligibility.

We ask that the Standing Committee carefully consider these concerns as they evaluate the measure.

National Quality Forum. Improving Attribution Models. Final Report. August 31, 2018. Available at: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=88154. Last accessed December 18, 2018.

National Quality Forum. Evaluation of the NQF Trial period for Risk Adjustment for Social Risk Factors. Final report. July 18, 2017. Available at: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=85635. Last accessed December 18, 2018.

Combined Methods Panel Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3612

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

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☒ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

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

2. **Briefly summarize any concerns about the measure specifications.**

   Panel Member 3: The outcome attribution process as defined is very complex, it isn't clear that visits used for attribution to a PCP or cardiologist are for heart failure, the definition of "dominance" of a provider, based solely on visit frequency vs. visit acuity is not conceptually or empirically supported, and the probability of attribution of roughly equal numbers of visits for patients with multiple visits (high users) to both PCPs and specialists is not mentioned. The stability of assignment to an individual provider over the time interval is also not addressed, nor is the assignment to multiple clinicians within the same group.

   Panel Member 4: The following exclusions are not defined in the testing form nor the XL data dictionary file: [1] 'Patients who were in hospice…' [2] 'Patients who had no E&M visits…'

   Panel Member 7: none

   Panel Member 8: It is not clearly stated that hospital admissions will be based on a heart failure diagnosis as the primary diagnosis. Patients with heart failure may require admission for many unrelated reasons

but will almost always have heart failure represented on their claims. Therefore, essential to include only patients for whom heart failure is the primary discharge diagnosis.

Panel Member 9: None

RELIABILITY: TESTING

Type of measure:

☒ **Outcome (including PRO-PM)** ☐ **Intermediate Clinical Outcome** ☐ **Process**

☐ **Structure** ☐ **Composite** ☐ **Cost/Resource Use** ☐ **Efficiency**

Data Source:

☐ **Abstracted from Paper Records** ☒ **Claims** ☐ **Registry**

☐ **Abstracted from Electronic Health Record (EHR)** ☐ **eMeasure (HQMF) implemented in EHRs**

☐ **Instrument-Based Data** ☒ **Enrollment Data** ☒ **Other (please specify)**

Panel Member 5: MIPS eligible provider files for 2017 and 2018

Panel Member 6: AHRQ SES index score

Level of Analysis:

☒ **Individual Clinician** ☒ **Group/Practice** ☐ **Hospital/Facility/Agency** ☐ **Health Plan**

☐ **Population: Regional, State, Community, County or City** ☐ **Accountable Care Organization**

☐ **Integrated Delivery System** ☐ **Other (please specify)**

Measure is:

☒ **New** ☐ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

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

3. **Reliability testing level** ☒ **Measure score** ☐ **Data element** ☐ **Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure**
   ☒ **Yes** ☐ **No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of **patient-level data** conducted?
   ☐ **Yes** ☐ **No**
6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   Panel Member 1: Signal-to-noise ratio from hierarchical GLM

   Panel Member 2: The ICC formula used by the developers is not a formula I recognized and the full technical details were not given within the measure testing document. Could the measure developers clarify what the ICC represents in this context? Can the reported reliability estimates be interpreted as squared correlations between estimated and true values? Some other interpretation?

   Panel Member 3: The developer used signal-to-noise analysis to assess reliability based on between clinician/group variance generated by HLM. They estimated the sample size required to achieve a minimum reliability of 0.40 using what appears to be a variant of the Spearman Browne Prophecy formula.

   Panel Member 4: The type of test was appropriate. However, 4 issues: [1] The measure steward submitted the measure for endorsement for: a) groups, b) individual clinicians. There are select tests at the group level, but it's unclear which (if any) tests were conducted at the individual clinician level. [2] Selective SNR

testing results were computed & reported, specifically: [a] percent of groups/individuals with a volume of 21 or greater with an R equal to (but I believe they meant equal to or greater) than 0.4. [b] percent of groups/individuals with a volume of 32 or greater with an R equal to (but I believe they meant equal to or greater) than 05. [3] Table on p. 8 under "distribution of reliability scores" reports out test results from "maximum" to "minimum" & several gradations in between. However, they fail to define what unit of analysis this is in regard to. [4] The test result reported in 2a2.4 fails to define what unit of analysis this is in regard to.

Panel Member 5: median signal-to-noise reliability for all clinician groups was 0.183 A minimum reliability of 0.4 was achieved for TINs with at least 21 HF patients. At this threshold, reliability scores for TINs ranged from 0.40 to nearly 1.0, with a median value of 0.600 (IQR 0.481-0.778). With the 21-patient volume minimum, the measure included 23.9% of clinician groups; however, 88.9% of the patients, 91.3% of the admissions, and 69.8% of clinicians, who reported under these TINs. This is acceptable reliability for TINS w/> 20 patients

Panel Member 6: No major concerns

Panel Member 7: Signal-to-noise reliability estimated using Nakagawa's formula.

Panel Member 8: Combination of signal to noise and ICC

Panel Member 9: Reliability testing methods were appropriate.

7. **Assess the results of reliability testing**

Submission document: Testing attachment, section 2a2.3

Panel Member 1: With at least 21 patients in the TIN, median reliability was 0.6. With at least 32 patients in the TIN, median reliability was 0.7. Patients attributed to TINs with at least 32 patients constituted 85% of all qualifying patients.

Panel Member 2: The median estimated reliability score was 0.183 which sounds quite low. Reliabilities were much higher (median = 0.60) among providers with at least 21 eligible cases. Only 24% of providers had this many cases which again sounds a bit low. However, in absolute terms, there were >10,000 providers with at least 21 eligible cases. This suggests that the measure could be usefully applied to a large number of providers.

Panel Member 3: The average reliability of across all clinicians/groups was poor (0.183). Limiting the sample to the clinicians/groups with volumes ≥21 pts (23.9% of the original sample) resulted in a higher reliability coefficient (0.40). Further limiting the sample to those clinicians/groups with ≥32 patients (16.8% of the original sample) resulted in a reliability coefficient of 0.50. The distribution of the sample also appears to favor group practices, but it is not clear whether these are small/large group practices or what specialties are represented in these practices. Reliability coefficients even at the higher level reported would still be of concern for between group comparisons. It is also not clear whether the methods used to estimate reliability have taken provider/patient ratios into account, i.e., whether higher volume practices have a large number of clinicians each seeing a small number of patients vs. individual providers with a high volume of patients. The HLM procedures as reported may or may not take this structure into account.

Panel Member 4: Regarding group level test results: [1] 23.9% of groups had an R value of 0.4 or greater where the minimum n was 21 [Table, p8] [2] 16.8% of groups had an R value of 0.5 or greater where the minimum n was 32 [Table, p8] [3] Median R value for groups with a minimum n of 21 was 0.6 Thus, a "moderate" rating for groups. Regarding individual clinician level test results: No individual level clinician test results calculated. Thus, an "insufficient" rating for individual clinicians.

Panel Member 6: I am a bit concerned that to reach an acceptable level of reliability, the developer only included 23.9% of provider groups and 69.8% of clinicians in the testing sample. This seems like a bias to me.

Panel Member 7: With recommended volume threshold of 21, median ICC=0.60 and minimum ICC=0.40.

Panel Member 8: Signal to noise 0.183 If restricting to providers with at least 21 heart failure patients, median of 0.6, but this represents only 23.9% of clinical groups and only 69.8% of clinicians belonging to those groups even though it represents 88.9% of patients and 91.3% of admissions (but assessment is at the clinician level)

Panel Member 9: Reliability is adequate above sample sizes of at least 21 patients per individual or practice being evaluated.

8. 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.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☒ **No**

   ☐ **Not applicable** (score-level testing was not performed)

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

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☐ **No**

   ☒ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications 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** (NOTE: Should rate **LOW** if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate **INSUFFICIENT** if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

    Panel Member 1: Moderate reliability observed among majority of attribution units (generally, practices)

    Panel Member 3: Both concerns about attribution specifications and the relatively low reliability levels, even for higher volume practices

    Panel Member 4: The following exclusions are not defined in the testing form nor the XL data dictionary file: [1] 'Patients who were in hospice…' [2] 'Patients who had no E&M visits…' [response to Q2] Regarding group level test results: [1] 23.9% of groups had an R value of 0.4 or greater where the minimum n was 21 [Table, p8] [2] 16.8% of groups had an R value of 0.5 or greater where the minimum n was 32 [Table, p8] [3] Median R value for groups with a minimum n of 21 was 0.6 Thus, a "moderate" rating for groups would have been given if the technical specifications were adequately defined. Regarding individual clinician level test results: No individual level clinician test results calculated. Thus, an "insufficient" rating for individual clinicians would have been given if the technical specifications were adequately defined...

    Panel Member 5: median signal-to-noise reliability for all clinician groups was 0.183 A minimum reliability of 0.4 was achieved for TINs with at least 21 HF patients. At this threshold, reliability scores for TINs ranged from 0.40 to nearly 1.0, with a median value of 0.600 (IQR 0.481-0.778). With the 21-patient volume minimum, the measure included 23.9% of clinician groups; however, 88.9% of the patients, 91.3%

of the admissions, and 69.8% of clinicians, who reported under these TINs. This is acceptable reliability for TINS w/> 20 patients

Panel Member 6: See comments above

Panel Member 7: Contingent on recommended volume threshold.

Panel Member 8: extremely low reliability for large portion of those entities being measured

Panel Member 9: Reliability is in the .6 range for entities with sample sizes greater than 21. Many individuals or practices in MIPS have fewer patients in the denominator than this, so an endorsement should specify that the measure should only be used with denominators at 21 or greater.

VALIDITY: TESTING

12. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**
13. **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.

    Submission document: *Testing attachment, section 2b1.*

    ☐ **Yes**

    ☐ **No**

    ☒ **Not applicable** (data element testing was not performed)
14. **Method of establishing validity of the measure score:**

    ☒ **Face validity**

    ☒ **Empirical validity testing of the measure score**

    ☐ **N/A (score-level testing not conducted)**
15. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

    ☒ **Not applicable** (score-level testing was not performed)
16. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

    Panel Member 1: Expert review

    Panel Member 2: Results in 2b1 focused on face validity as determined by an expert panel. Empirical analyses of risk model performance are also relevant to the assessment of validity. These were presented in section 2b3.

    Panel Member 3: The TEP provided survey based assessments of face validity.

    Panel Member 4: Face validity was used where the panelists were asked 2 questions. The 2 questions were in regard to perceptions of the ability for the measure to discern quality at the group level. Given the measure was submitted at the group & individual clinician level: [a] the testing was adequate for groups as the unit of analysis [b] the testing was insufficient for individual clinicians as the unit of analysis

    Panel Member 7: Technical Expert Panel and Clinician Committee.

    Panel Member 8: Majority of expert panel agreed that measure could distinguish higher quality providers and provide information related to quality

    Panel Member 9: For a new measure, face validity is fine, and there is a formal method for establishing face validity here.

17. **Assess the results(s) for establishing validity**

Submission document: Testing attachment, section 2b2.3

Panel Member 1: Majority of experts felt that that measure provides useful information. Not especially persuasive.

Panel Member 2: The TEP appeared to have a lukewarm assessment of the measure's utility. The risk model appeared to be suitable to adjust for case mix.

Panel Member 3: At this phase of measure development, face validity alone appears adequate.

Panel Member 4: The face validity results at the group level were moderate given the groups' voting results for the 2 questions. Given the 2 questions were regarding the group as the unit of analysis, there are no / insufficient findings as the individual clinician unit of analysis.

Panel Member 7: Only 12 of 17 active TEP members voted - why? TEP and Clinician Committees generally supported value of the measure, although with a few dissenters.

Panel Member 8: Admission is a frequent necessity for heart failure patients. Lack of admission may reflect poor access to care or even poor care. Taken in vacuum without correlation to other important metrics such as mortality, highly questionable validity

Panel Member 9: Face validity is adequate - moderate.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

18. **Please describe any concerns you have with measure exclusions.**

Submission document: Testing attachment, section 2b2.

Panel Member 1: Why CKD-5 is excluded because of nephrology care, but CKD-4 is not? CKD-4 is, by guideline, under nephrology care, not cardiology care.

Panel Member 4: No concerns other than the fact some are not defined, which is discussed in response to Q2.

Panel Member 7: None, all exclusions are reasonable

Panel Member 9: None

19. **Risk Adjustment**

Submission Document: Testing attachment, section 2b3

19a. **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

19b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☒ Yes ☐ No ☒ Not applicable

19c. **Social risk adjustment:**

19c.1 Are social risk factors included in risk model? ☒ Yes ☒ No ☐ Not applicable

19c.2 Conceptual rationale for social risk factors included? ☒ Yes ☐ No

19c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☐ No

19d.**Risk adjustment summary:**

19d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

19d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

19d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

19d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☐ No

19d.5.Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☐ No

19e**. Assess the risk-adjustment approach**

Panel Member 1: Hierarchical GLM with adjustment for age, comorbidity, and low SES index score

Panel Member 3: Approach appears adequate and includes AHRQ SES Index, however the developers observed that the later did not significantly contribute to the explained variance in admission rates.

Panel Member 4: The risk adjustment strategy is appropriate given the type of measure and population measured. The tests conducted to evaluate the risk model were adequate. The findings from the risk model testing suggest the risk model is sufficient.

Panel Member 7: Risk-adjustment model lacks indicators of heart failure severity, so it seems unlikely to be able to account for differences in case mix between primary care physicians and specialty cardiologists. However, inclusion of functional and frailty markers is commendable. Overall model performance is probably adequate at R2=0.072 with good calibration.

Panel Member 8: Model does not account for the repeated measures impact--i.e., for impact of single patient with multiple admissions vs. multiple patients with single admission. A single problematic patient may make individual provider look like delivering lower quality care--perhaps why there is such low reliability. Race is not accounted for or even tested in the modeling. Response to various medications has been shown to differ by race. Exclusion of this factor based on policy rather than data and science is problematic. Model accounts for 7.3% of the variance but c-statistic of the model is not given. Calibration appears to be adequate.

Panel Member 9: The logic here seems a little strained, as it results in the AHRQ SES variable being included but dual-eligible status not included, even though there seems to be some evidence in favor of both.

20. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

    Submission document: Testing attachment, section 2b4.

    Panel Member 3: There appears to be significant clinician/group level variation in the readmission rates. It would be informative to provide the standard error of measurement by clinician/group physician and patient size.

    Panel Member 4: Issues with the testing in this regard: [1] Results presented were expressed in percentiles. I could not consider this as responsive to the question to identify "meaningful" differences. [2] Results presented were at the group level. Again, the measure submitted for endorsement was at the group level and the individual clinician level. Thus, the measure steward should have also presented findings at the individual clinician level as well.

    Panel Member 7: none

    Panel Member 8: See answer to #19

    Panel Member 9: No way to define what a meaningful difference is.

21. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

    Submission document: Testing attachment, section 2b5.

    Panel Member 4: NA – multiple methods were not used.

    Panel Member 7: none

    Panel Member 9: N/A

22. **Please describe any concerns you have regarding missing data.**

    Submission document: Testing attachment, section 2b6.

    Panel Member 4: No concerns.

Panel Member 7: none

Panel Member 8: no concerns

Panel Member 9: None

For cost/resource use measures ONLY:

23. **Are the specifications in alignment with the stated measure intent?**

    ☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

24. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

25. **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** (NOTE: Should rate LOW if you believe that there **are** threats to validity and/or relevant threats to validity were **not assessed OR** if testing methods/results are not adequate)

    ☒ **Insufficient** (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.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

    Panel Member 1: R^2 = 0.073 is really quite good for a hospitalization rate model

    Panel Member 2: The TEP's impression of the measure's validity was moderate to low

    Panel Member 3: Although face validity appears adequate, no empirical testing of validity was performed.

    Panel Member 4: The face validity results at the group level were moderate given the groups' voting results for the 2 questions. Thus, the 'moderate' response to Q25 is regard the group level unit of analysis. Given the 2 questions were regarding the group as the unit of analysis, there are no / insufficient findings at the individual clinician unit of analysis. Thus, evaluating validity at the clinician level is not possible given the testing results are missing for this unit of analysis.

    Panel Member 5: model exhibited acceptable calibration and discrimination in the validation data: -The deviance R-squared for the model with demographic and clinical risk factors was 0.073 in the Development HF Full Sample and 0.072 in the Validation HF Full Sample -in the Validation Sample, the over-fitting index of γ0 was close to 0 (-0.007) and γ1 was close to 1 (0.993)

    Panel Member 8: Failure to explore race in risk model, failure to account for repeated measures effect of problem patients and unknown c-statistic of risk model

    Panel Member 9: Face validity only for this measure, and it is marginally acceptable.

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?**

    ☐ **High**

    ☐ **Moderate**

☐ **Low**

☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

    Panel Member 8: As noted above, this measure could theoretically penalize a careful provider who is careful to provide in-patient care to a very sick patient population