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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 2321

Corresponding Measures:

De.2. Measure Title: Functional Change: Change in Mobility Score

Co.1.1. Measure Steward: Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.

De.3. Brief Description of Measure: Change in Rasch derived values of mobility function from admission to discharge among adults aged 18 and older receiving inpatient medical rehabilitation at a post-acute care facility who were discharged alive. The timeframe for the measure is 12 months. The measure includes the following 4 mobility items:1. Transfer Bed/Chair/Wheelchair, 2. Transfer Toilet, 3. Locomotion, 4. Stairs.

1b.1. Developer Rationale: The current mandated quality measures for inpatient rehabilitation facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of inpatient rehabilitation is to increase patient function to return the patient to home/previous residence within the community. Yet the current measures don't adequately capture function or functional improvement. The current quality indicator measures address facility level process, which, has been argued, is not applicable to the inpatient medical rehabilitation setting as the overall prevalence of these events are very low (less than 2% of patients affected per year) and often times, the presence of the quality indicator occurred in the acute care setting or prior to admission to acute or post-acute care (for instance, CAUTIs and incidence of new or worsened pressure ulcers).

The mobility measure is constructed by utilizing items from the FIM® instrument, which is presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the FIM® instrument. The FIM® instrument is already used in inpatient rehabilitation as it is embedded in the IRF-PAI, which is required to be completed for payment reimbursement by CMS. Each of the four items that comprise the mobility measure are presently collected in the IRF-PAI to capture patient functional status. Utilizing the change in mobility function measure as a quality indicator would not create any additional costs to IRFs, since IRFs are already transmitting the current IRF-PAI data to CMS for payment purposes. The change in mobility measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and is predictive of patient change in mobility function outcomes and likelihood of patient discharge from inpatient rehabilitation to home/the

community. It is imperative that any quality indicators used in the post-acute care setting take into account the overriding goal of medical rehabilitation, which is to restore and improve patient function and increase functional independence among individuals thus allowing the patient the ability to return to a community setting upon discharge from an inpatient facility.

S.4. Numerator Statement: Average change in Rasch derived mobility function score from admission to discharge at the facility level. Includes the following items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. Average is calculated as: (sum of change at the patient level/total number of patients). Patient less than 18 years of age at admission to the facility or patients who died within the facility are excluded.

S.6. Denominator Statement: Facility adjusted expected change in Rasch derived mobility values, adjusted at the Case Mix Group (CMG) level.

S.8. Denominator Exclusions: National values used in the CMG adjustment procedure will not include cases who died in the IRF or patients less than 18 years of age at admission. Cases who died during rehabilitation are not typical patients and are routinely omitted from reports and published research on rehabilitation outcomes. Further details and references related to the exclusion criteria can be found in the Measure Testing form.

De.1. Measure Type: Outcome

S.17. Data Source: Instrument-Based Data, Other

S.20. Level of Analysis: Facility, Other

IF Endorsement Maintenance – Original Endorsement Date: Nov 04, 2015 **Most Recent Endorsement Date:** Nov 04, 2015

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

Preliminary Analysis: Maintenance of Endorsement

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

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 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

- Brief background: this is a measure of functional status change assessing four different mobility functions for patients 18+ as assessed by a clinician. The items that comprise the mobility measure include: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
- The primary aim of inpatient rehabilitation is to restore function, increase functional independence, and to discharge the patient back to home/the community setting or residence prior to the patient's acute admission and/or IRF stay.
- The measure is informed by the FIM instrument, a tool used in inpatient medical rehabilitation to assess the patient's level of functional status at admission and at discharge. The FIM instrument includes 18 items, of which, four items address patient mobility function.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- Developer provides a logic model depicting the relationship between IRF admission, patient goals and rehab plan, treatment, discharge, and mobility change in function.
- Developer did not provide a healthcare structure, process, intervention, or service that can be deployed to improve performance on the measure. Presumably, the provision of rehabilitation services would lead to improvement in scores, but the developer did not offer evidence of this.
- Developer instead offered evidence that the self-care measure correlates to positive outcomes
    - Measure 2321 correlates statistically significantly to the FIM instrument (p<.001). The measure's mobility care correlation:
        - Admission mobility correlation to the admission FIM motor total was .82
        - Discharge mobility correlation to the discharge FIM motor total was .93
        - Total change mobility correlation to the total change FIM motor score was .87
- The mobility care measure, independently, was assessed to determine predictive ability of the measure on patient outcomes. All 4 items of the mobility measure were retained in each of the regression models and were statistically significant (p<.001) in the models.
    - Significant predictor of patient discharge to the community, chi-square=46078.9, (df=4), p<.001, R2 =.14.
    - Significant predictor of patient LOS, adjusted R2 =.15, p<.001.
    - Significant predictor of patient functional change from admission to discharge, adjusted R2= .27, p<.001.

Question for the Committee:

- *The developer did not provide empirical evidence of a structure, process, intervention or service that can improve performance. Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm

Measure assesses outcome (box 1) YES -> relationship between outcome and at least one healthcare action (box 2) NO -> NO PASS

Preliminary rating for evidence: ☐ Pass ☒ No Pass

RATIONALE:

NQF's Evidence requirements indicate a developer must provide empirical evidence that there is at least one structure, process, intervention, or service that can improve performance. Therefore the preliminary analysis is rated as No Pass. However, if the developer can provide this evidence at the Committee measure evaluation meeting, and the Committee agrees it meets the criteria, the measure would be eligible for a 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.

- Developer analyzed mean change in mobility scores at the facility level and then grouped facilities by performance quartile (not quartiles based on number of facilities).
- Quartile 1 (25th%): (n = 10), Mean - 2.8, Standard Deviation - 2.6
- Quartile 2 (25th-50th%): (n = 538), Mean - 8.6, Standard Deviation - 1.1
- Quartile 3 (50th-75th%): (n = 197), Mean - 11.5, Standard Deviation - 0.5
- Quartile 4 (75th%): (n = 5), Mean - 15.2, Standard Deviation - 2.0
- This suggests a fairly narrow performance range with the overwhelming majority of facilities within a few points of each other.

Disparities

- The developer assessed disparities in performance for the following social risk factors: race, sex, and marital status. Across all three groups assessed, no differences in mean change in mobility score were evident. The change in total mobility scored from admission to discharge by group was:
  - Race (all race and ethnic categories) - eta2 < .001
  - Sex - eta2 < .001
  - Marital status - eta2 < .001

Questions for the Committee:

- Is there a meaningful spread of performance between the entities evaluated by the measure developer?
- Is measure gap sufficient for the Committee to assume that there may be actionable differences and not just differences in population that account for the measure performance gaps, and thus represent evidence of structures, processes, interventions or services that improve performance?

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

RATIONALE

- Measure demonstrates a narrow range of performance by the overwhelming majority of facilities.
- This is indicative of poor opportunity for improvement and warrants consideration by the Committee for a low rating.

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

1a. Evidence: *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.

- The FY 2019 Medicare Inpatient Rehabilitation Facility Prospective Payment System Final Rule (CMS-1688-F) states that CMS is removing the FIM instrument and associated Function Modifiers from the IRF-PAI for discharges beginning on or after October 1, 2019. This is being done to reduce the burden on IRFs as the FIM measures for mobility will now be measured by the IRF quality metrics known as Section GG. CMS found the FIM to be more burdensome than informative and noted that the entities were all clustered at the top of the range, reducing its ability to distinguish one from another.
- Unclear what interventions will move scores and whether facilities and implement changes to improve outcomes

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?*

- Again, performance gap is somewhat narrow, limiting its utility to distinguish one entity from another.
- Performance gap was quite narrow except for outliers at either extreme. Does not appear to be sensitive. Disparities showed no difference by sex, race or marital status. No clear relationship between performance and specific interventions.

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.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators: NQF Scientific Methods Panel

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

Scientific Methods Panel Votes: Measure Passes

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

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

- Testing included score-level and data element testing
- The developer conducted reliability testing for both data element and measure score. For data element reliability, the developer reported internal consistency. For measure score reliability, the developer conducted split-half reliability testing based on a random sample of facilities.
- SMP considered both testing methods to be appropriate, however, it was not obvious why split-half reliability testing could not be conducted with all facilities: "At minimum, it will be useful for the developer to provide additional descriptive information of the random sample. Volume information is particularly relevant, as it will directly impact reliability estimated."
- SMP assessment of results:
    - Alpha (0.78) is good, but some concern over low item-total correlations for memory and dressing-lower extremity
    - Score level reliability across facilities was quite good (ICC=0.95)
- SMP notes/comments to measure developer:
    - Why was it necessary to do a random sample of 30 facilities instead of using the full set of 855 facilities? This would have been more informative.
    - "A stronger method of reliability testing would include an analysis of within- facility score and between-facility score variation to understand the strength of the 'signal' represented in measure scores. Alternatively, adding bootstrapping to the ICC analysis would make the analysis more robust."

Validity

- Testing included score-level and data element testing
- Measure developer employed construct validity, predictive validity for the item level analysis and criterion-referenced validity for the score-level. The developer evaluated construct, criterion and predictive validity. Criterion validity focused on the relationship between the self-care items and the full FIM instrument, whereas predictive validity focused on patient outcomes (discharge to the community, LOS). The SMP considered the predictive validity analyses appropriate and compelling.

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**

Measure Number: 2321

Measure Title: Functional Change: Change in Mobility Score

Type of measure:

☒ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☐☐ **Outcome** ☒☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☐ **Claims** ☒☐ **Electronic Health Data** ☒☐ **Electronic Health Records** ☐ **Management Data**

☐ **Assessment Data** ☐ **Paper Medical Records** ☐☒ **Instrument-Based Data** ☒☐ **Registry Data**

☐ **Enrollment Data** ☒☐ **Other:**

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☒ **Facility** ☐ **Health Plan**

☐ **Population: Community, County or City** ☐ **Population: Regional and State**

☐ **Integrated Delivery System** ☒☐ **Other** patient level change in function, including **aggregate**

Measure is:

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

MP#2: Average change in Rasch-scored FIM score from admission to discharge. 4 mobility items from transfer to toilet to locomotion.

Data source: Patient data from the FIM® instrument collected from inpatient rehabilitation facilities, long term acute care facilities, and skilled nursing facilities subscribing to the Uniform Data System for Medical Rehabilitation (UDSMR).

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.**

   MP#1: Specs are not clear but this measure has been implemented for years and the measure is not in the public domain, so this may be less of an issue.

   MP#6: None

RELIABILITY: TESTING

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** NA

6. **Assess the method(s) used for reliability testing**

   MP#6: Rasch analysisappears to be an appropriate conversion for the analysis along with the Cronbach's alpha. ICC split half method for the facility analysis did not raise concerns.

   MP#2:

   a. Methods were appropriate. Cronbach's alpha and Rasch analysis for 4-item FIM, ICCs for comparison of reliabilities across facilities

   Submission document: Testing attachment, section 2a2.2

   MP#4: Appropriate

   MP#3: The developer conducted reliability testing for both data element and measure score. For data element reliability, the developer reported internal consistency. For measure score reliability, the developer conducted split-half reliability testing based on a random sample of facilities. Both testing methods were appropriate, however, it is not obvious why split-half reliability testing could not be conducted with all facilities. At minimum, it will be useful for the developer to provide additional descriptive information of the random sample. Volume information is particularly relevant, as it will directly impact reliability estimated.

   MP#1: Cronbach's alpha (data element) and ICC (measure score) were used to evaluate reliability. However, based on the results provided, it does not appear that an evaluation of between and within facility score variation performed (i.e. signal-to-noise).

   MP#5: The methods used were standard and generally acceptable. Cronbach's alpha was used to assess data element (instrument) reliability, and a split-half ICC test was conducted for reliability at the measure score level. It is not clear at all why a random sample of 30 facilities was needed for the ICC analysis, as no other measure developers using this method seem to have a problem with a large number of facilities in a data base. Analysis on the full set of 855 facilities would have been more informative, although the random sample results are worth something.

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

   MP#6: All results supported reliability at the item and facility level.

   MP#2:

   a. Alpha (0.83) is good, but some concern over low item-total correlations for memory and dressing-lower extremity

   b. Score level reliability across facilities was quite good (ICC=0.95)

   Submission document: Testing attachment, section 2a2.3

   MP#4: No issues.

   MP#3: Cronbach's alpha was very good at 0.78.

   Measure score reliabliyt measured by ICC was excellent at 0.95.

   MP#1: Internal consistency and ICC facility-level correlations were results were positive however, this analysis does not provide strong evidence regarding the precision of the measure score results (only that scores are correlated across facilities). A stronger method of reliability testing would include an analysis of within- facility score and between-facility score variation to understand the strength of the 'signal' represented in risk adjusted measure scores.

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.**

    MP#3: Comprehensive reliability tests were conducted, covering both data element and measure score. The results were very good. Moderate rating is partly due to split-half relability testing was only based on a small sample of facilities.

    MP#4: Based on methods used and testing results.

    MP#6: No concerns, analysis at the item and facility levels both supported reliability tests were conducted, covering both data element and of the measure.

    MP#1: A stronger method of reliability testing would include an analysis of within- facility score. The results were and between-facility score variation to understand the strength of the 'signal' represented in measure scores. Alternatively, adding bootstrapping to the ICC analysis would make the analysis more robust.

    MP#2:

    a. Very good internal consistency and score-level reliability for a well-defined construct. No missing data reported

    MP#5: As noted above, the statistical results of reliability testing look very good, but the use of a small random sample of facilities for the ICC analysis was strange.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    MP#2: Death and under 18 exclusion ok and well-defended

    MP#6: No concerns

    Submission document: Testing attachment, section 2b2.

    MP#4: No concerns.

    MP#3: No concern

MP#5: None.

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

    MP#2: Highly-significant F statistic in comparing 4 quartiles on score change values

    MP#6: Mean change in mobility scores at the facility level were computed and mobility change scores were grouped by quartile. ANOVA analysis indicates statistical differences in mean change scores by quartile

    Submission document: Testing attachment, section 2b4.

    MP#4: No concerns.

    MP#3: No concern.

    MP#1: None.

    MP#5: The developers could show that the facilities could be arrayed from high to low (or vice-versa) in terms of measure score, and quartiles established based on that distribution. There are differences in score among the quartiles. This speaks to the issue of there being differences at all among facilities, but does not bear on the question of whether the differences are meaningful. That analysis would require some link to information on clinical significance to patients or caregivers of a given difference in the FIM score.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    Submission document: Testing attachment, section 2b5.
    MP#4: NA
    MP#3: No concern.
    MP#1: N/A
    MP#5: None

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

    MP#6: Missing data was eliminated by sampling method

    MP#2:

    a. Submitters report there is no missing data. with such a huge data source, this is hard to believe. Perhaps this is linked to the definition of denominator? Surely there is missing FIM data in IRFs? I realize they get all data on IRF discharges, but do all have FIM scores?

    Submission document: Testing attachment, section 2b6.

    MP#4: None.

    MP#3: No concern

    MP#5: None.

16. **Risk Adjustment**

    16a. **Risk-adjustment method** ☐ **None** ☐☒ **Statistical model** ☐☒ **Stratification**

    16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

    ☐ Yes ☐ No ☒☐ Not applicable

    16c. **Social risk adjustment:**

    16c.1 Are social risk factors included in risk model? ☒☐ Yes ☐☒ No ☐ Not applicable

    16c.2 Conceptual rationale for social risk factors included? ☐☒ Yes ☒☐ No

    16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐☒ Yes ☒☐ No

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion?

☒☐ Yes ☐ No NA

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration)

MP#6: results indicated there were no differences in mobility score by race, sex or marital status

☒☐ Yes ☐☒ No

16d.5.Appropriate risk-adjustment strategy included in the measure? ☒☐ Yes ☐☒ No

16e**. Assess the risk-adjustment approach**

MP#4: Appropriate

MP#6: Case Mix Group (CMG) through indirect standardization is appropriate and demonstrated statistical differences between impairment types.

MP#3: Overall, the risk adjustment approach was acceptable.

MP#1: Difficult to assess the approach without information about the performance of the model. The developer provides mean change in mobility by CMG group. Model lacks social risk factors but developer notes mobility scores did not vary by race, sex or marital status. CMG is a reasonable approach but lacks important social risk factors.

MP#2: Uses CMS case-mix group specifications. With functional outcomes of inpatient care one must control for differences in patient impairment types/conditions and for the severity within a given impairment type/condition. Stratification for patient impairment type/condition and risk adjusting data by CMG has been used extensively in prior, published research on patient functional outcomes of inpatient rehabilitation.

A statistically significant difference was found in mean mobility change by impairment type, F=1021.40 (df=15), p=.000. Stroke was on the low end of change compared to other conditions.

MP#5: I think the risk adjustment method is probably better than the results and description provided in the testing form**.** There are apparently adjustments made with a range of clinical and demographic variables, but the testing form only describes one example of stratification in any detail. The example given is stroke, implying not only stratification, but a selection of the subset of patients with stroke for measurement and analysis. While this is a perfectly acceptable example of use of the measure, the measure is not specified or tested for reliability for stroke only, so there is a "disconnect" between the reliability data and the information provided on risk adjustment. What should have been provided would be data on the full adjustment or standardization model for calculation of the scores at the facility level.

<u>For cost/resource use measures ONLY</u>:

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

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

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

VALIDITY: TESTING

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

20. **Method of establishing validity of the measure score:**

☒☐ Face validity

☒☐ Empirical validity testing of the measure score

☐☒ N/A (score-level testing not conducted)

21. **Assess the method(s) for establishing validity**

MP#6: Construct validity, predictive validity for the item level analysis and criterion-referenced validity for the score-level were appropriate.

MP#2:

a. To calculate the facility's adjusted expected change in Rasch derived values, indirect standardization was used which weights national CMG-specific values by facility-specific CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The CMG classification system groups similarly impaired patients based on functional status at admission or patient severity. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes.

Submission document: Testing attachment, section 2b2.2

MP#4: Appropriate

MP#3: Extensive validity tests were conducted, including construct validity, predictive validity, criterion validity, and others. All testing methods were appropriate.

MP#1: The developer evaluated construct, criterion and predictive validity. Criterion validity focused on the relationship between the mobility item and the full FIM instrument, whereas predictive validity focused on patient outcomes (discharge to the community, LOS). The predictive validity analyses were appropriate and compelling.

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

MP#6: Regression models for predictive validity demonstrated significant results for DC to community, LOS and discharge disposition (home vs acute care facility). Stepwise regreassion demonstrated all items were predictive ability.

MP#2:

a. the full set of CMG analyses were not included in the submission. To illustrate the risk adjustment by CMG, stroke (the most common impairment) was presented and the results on FIM average change by CMG group were as predicted/expected.

Submission document: Testing attachment, section 2b2.3

MP#4: Appropriate

MP#3: Factor analysis identified one meaningful component that accounts for substantial total variance. All I tems were predictive of several relevant dependent variables.

The results of the criterion-referenced validity testing indicated a very strong correlation between the mobility measure and the FIM Instrument.

MP#1: The predictive validity analysis showed a strong and significant relationship between the mobility item and outcomes (discharge to the community, LOS, change in function), providing evidence of the validity of the measures.

MP#5: As far as I can tell, the testing for validity was done only at the data element (instrument) level. I see no testing of measure score validity. The data element validity testing was reasonable and used multiple analytic methods.

23. **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)

24. **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)

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.**

    MP#4: Testing results . Stron correlations with the FIM instrument.

    MP#6: As noted above, all analysis supported validity of both the item level and facility level scores

    MP#3: Testing of both measure score and data element validity showed good results, particularly measure score validity.

    MP#1: Very good predictive validity results, would have been helpful to see the full model results as well as to include other covariates included in the multivariable model of discharge to community and LOS to understand the unique contribution provided by the mobility item and other covariates.

    MP#2: Rating for stroke subgroup is moderate. Not seeing the other data, although I suspect it is probably ok, leads me to say insufficient with regard to a broad PM across conditions

    MP#5: As noted above, I see no data on validity at the measure score level. Since the emasure is based on an instrument (FIM), both types of validity testing are required.

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.**

Committee Pre-evaluation Comments:
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. 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?*

- From its inception the FIM has been a challenging instrument with regard to interrater reliability. So much so that the FIM required certification for staff to properly administer. Evidence of IRR was not provided. Reliability is very high when completed by certified staff and the logic and calculations are appropriate.
- all are well defined and consistent.

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

- No data on IRR offered.
- no, long experience and used in IRF-PAI for submission to CMS

2b2. Validity testing: *Do you have any concerns with the testing results?*

- No converns.
- No

Validity- Threats to Validity: *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?*

- No concerns in this area.
- 2b4: does not identify meaningful differences. The spread of scores is narrow and there is no corelation to outcomes or specific interventions. 2b5 NA 2b6: very little missing data (like due to secondary use in IRF-PAI)

Other Threats to Validity: *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?*

- The risk adjustment is based on CMS Case-Mix Group adjustments. Strategy and acceptability are good.
- No risks with one possible exception. There is a three day window for obtaining the admission FIM. From personal experience there is a difference in scores on day 1 vs day 3 (higher) meaning that the difference in dc and admission scores is higher if the measurement is done immediately on admission. WOuld be nice to see if this observation is born out and if so that this is adjusted for or constrained.

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.

- FIM tool data is collected by healthcare personnel during the provision of care and all data elements are defined fields in electronic clinical data

- The developer states the measure is publicly available for use free of charge. Facility-level and national benchmark reporting are available by the developer through a subscription; cost varies based on facility type and size.

Questions for the Committee:

- Does the Committee feel that the measure developer has provided sufficient information to determine how feasible the collection of this data is?

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?*

- Once CMS changes occur and the FIM is no longer a required tool none of the data elements will be routinely generated or used during care delivery. All required data elements will be according to Group GG quality metrics.
- High

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**

OR

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

Accountability program details

- CMS IRF-PAI (will no longer be required as of October 2019)
- Quality Improvement National IRF Benchmark Reports
- Quality Improvement Facility-level IRF Reports

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

- Feedback was not solicited from those being measured.
- Developer notes that the FIM instrument has been in use for over 25 years and required in IRF-PAI since 2002.

Additional Feedback:

Questions for the Committee:

Does the Committee have any concerns about the current or future use of this 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

- The developer did not offer year over year data to show improvement in self-care change scores amongst measured facilities over time.
- The developer instead showed that differences in average mobility change scores among differing facilities can be measured and rank ordered in terms of patient average change in mobility function from admission to discharge.
    - Statistically significant differences in mean change scores by quartile were determined, however standard deviation within quartiles were small.
    - Mean change scores and standard deviation by quartile:
        - Quartile 1 (25th%): Mean- 2.8, Standard Deviation- 2.6
        - Quartile 2 (25th-50th%): Mean- 8.6, Standard Deviation- 1.1
        - Quartile 3 (50th-75th%): Mean- 11.5, Standard Deviation- 0.5
        - Quartile 4 (75th%): Mean- 15.2, Standard Deviation- 2.0

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

- Statistically significant mean mobility differences arise by impairment type.
- The mobility measure can discriminate in-patient functional ability between different functional impairments and within the same type of functional impairment (ex. stroke).

Potential harms N/A

Additional Feedback: N/A

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

RATIONALE:

- Developer did not offer an evaluation of changes in performance over time.

- The submission is therefore insufficient for Usability.
- Note: this is not a must-pass criterion.

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

4a. 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?*

- Once fully implimented, the FY 2019 Final Rule for IRFs eliminates the collection of FIM mobility data and will no longer be publicly reported.
- Ok

4b. Usability: *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.*

- CMS has determined the burden outweighs the benefit. It has been retired due to the simplification of items for more precise and understandable data.
- No data to support quality as a function of score or outcomes as a function of scores. Need performance data over time by facility and all facilities that demonstrates improved scores AND improved outcomes (LOS, return to work, return home, days spent in facility)

Criterion 5: Related and Competing Measures

Related or competing measures

- This measure is competing with one measure: 2634: Inpatient Rehabilitation Facility (IRF) Functional Outcome Measure: Change in Mobility Score for Medical Rehabilitation Patients
- The Committee will need to compare both measures and attempt to reach a best-in-class decision. NQF staff will prepare additional materials to assist the Committee in this comparison.

Harmonization

Measure 2321 consists of four items rated on a 7-level scale, where clinicians rate the patient's lowest actual observed score for the past 24-hour period. Measure 2321 is similar to measure 2634 which includes 15 mobility items (measuring the same constructs, such as ambulation). Measure 2634 rates items on a 6-level scale and allows for options to not assess each item. By not assessing each item, the developer notes that the 6-level scale is less sensitive than the 7-level scale. The developer states this has potential to lead to determining the patient has "a higher level of function than truly exists". Additionally, the developer questions the validity of data interpretation based on the 6-level scale allowing for multiple missing options. "The inclusion of multiple ´missing´ options for each item to be allowed for use at admission and at discharge lends the possibility for data that is not able to be interpreted, if an item is not rated at admission because the patient refused but is rated at discharge, of what value is this information?" Measure 2321 is intended for all patients age 18 and older who receive post-acute care at an IRF, SNF, or LTAC facility, while measure 2634 is intended for Medicare patients who receive care an IRF.

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

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?*

- While the new CMS metrics were noted as competing, the developer did not offer any thoughts on harmonization.
- 2321, can't be harmonized (different scales, different items. Based on worksheet references the FIM significantly underestimates the need for assistance particularly around toileting. This is a significant failure of utility.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: June/13/2019

- As a family caregiver, I have been following the conversation re self-care/mobiity scores across the care continuum, including discharge to community. In the Fall 2017 report of the Patient Experience and Function Standing Committee, there appeared to be uncertainty aboutt the merits of Section GG vs FIMS. I've searched the Fall 2018 report and have had a hard time discerning whether of not this issue has been settled. In the event that the Standing Committee is still accepting comments on this issue, I urge that Section GG be selected. The Section GG 6pt scale clearly communicates the level to which a patient relies on personal assistance in a manner that the patient, clinician and family member can understand. Particularly in discharges to home, the family needs to appreciate the degree to which their loved one will be depending on their presence to perform self-care tasks.
- Please note: Study examined how similar summary scores of physical functioning using the Functional Independence Measure (FIM) can represent different patient clinical profiles. Data were analyzed for 765,441 Medicare fee-for-service beneficiaries discharged from inpatient rehabilitation. Patients' scores on items of the FIM were used to quantify their level of independence on both self-care and mobility domains. Patients requiring "no physical assistance" at discharge from inpatient rehabilitation were identified by using a rule and score-based approach. In patients with FIM self-care and mobility summary scores suggesting no physical assistance needed, the study found that physical assistance was in fact needed frequently in bathroom-related activities (e.g., continence, toilet and tub transfers, hygiene, clothes management) and with stairs. It was not uncommon for actual performance to be lower than what may be suggested by a summary score of those domains. The authors conclude that further research is needed to create clinically meaningful descriptions of summary scores from combined performances on individual items of physical functioning. Citation: Fisher, Steve R., Middleton, Addie, Graham, James E., Ottenbacher, Kenneth J.. (2018). Same but different: FIM summary scores may mask variability in physical functioning profiles. Archives of Physical Medicine and Rehabilitation , 99(8), Pgs. 1479-1482, 1482.e1. Retrieved 12/6/2018, from REHABDATA database.
- While this is not an eCQM, we would encourage the measure steward to use a standard terminology such as LOINC for encoding the FIM instrument in their measure. Without this level of standardization, interoperability will be a perpetual challenge, and impact the ability to measure a patietnt's functional status across the continumm of care.

1. Evidence and Performance Gap – 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 sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

NQF_evidence_attachment_2321_.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

Yes

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 2321

Measure Title: Functional Change: Change in Mobility Score

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

Date of Submission: 4/8/2019

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

Outcome

☒ Outcome: Functional Change; change in patient mobility score from admission to discharge

☐ Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:

☐ Appropriate use measure:

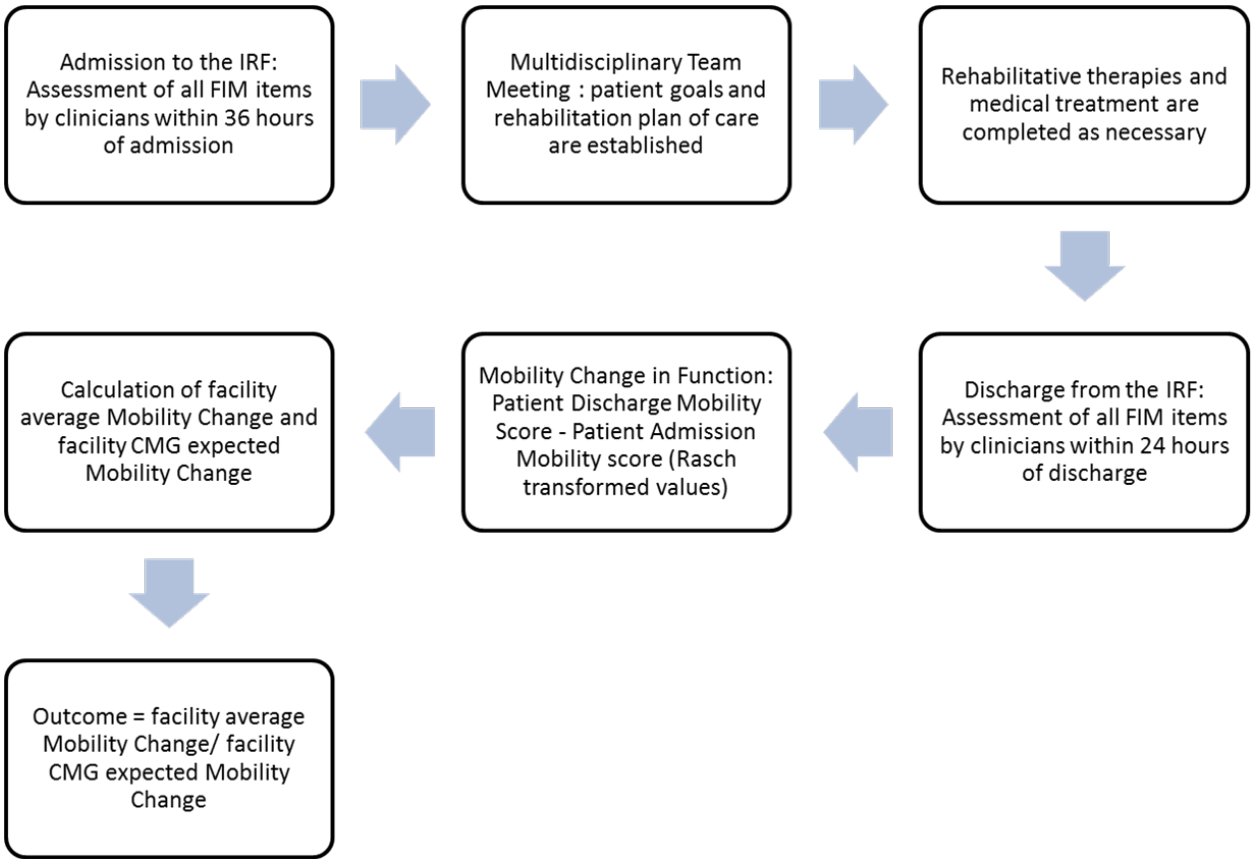
☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Inpatient rehabilitation facilities (IRFs) are one part of a multi-level post-acute care continuum. Inpatient rehabilitation is meant to provide intensive rehabilitation therapy for patients who, due to the complexity of their nursing, medical management and rehabilitation needs, require extensive rehabilitation therapy

utilizing a multidisciplinary team approach. The primary aim of inpatient rehabilitation is restore function, increase functional independence, and ideally, to discharge the patient back to home/the community setting or residence prior to the patient's acute admission and/or IRF stay. 1,2. The FIM instrument is presently embedded in CMS's IRF-PAI, which is the instrument used in inpatient medical rehabilitation to assess the patient's level of functional status at admission and at discharge. Completion of the IRF-PAI is required by CMS as part of prospective payment for facility reimbursement of services provided to the patient. The FIM instrument includes 18 items, of which, four items address patient mobility function. The mobility items have been extensively used for over 25 years as a component of the larger FIM instrument, in essence, the mobility measure is a measure within a larger measure. The mobility measure is to be administered within 24-36 hours of the patient's admission to the post acute facility and again on the day of patient discharge. Interim assessments can be performed for case management purposes (goal setting or altering the therapy) but are not required. The items that comprise the mobility measure include: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. All Items are assessed by trained clinicians at the post acute care facility. Below is a flow chart depicting the methodology for patient assessment of the measure:



1a.3 Value and Meaningfulness: **IF** this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

Mobility measure is not derived by patient report, it is clinician assessed.

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

As previously stated, the mobility measure items exist within a larger instrument, the FIM instrument, which has been widely used and extensively published upon. For these reasons, much of the rationale, feasibility,

usability and validity of the mobility measure is referenced to the larger FIM instrument, which is, in essence, the foundation. The FIM® instrument has been demonstrated in hundreds of peer-reviewed journal articles (see bibliography in Appendix) to be a significant predictor of patient outcomes of rehabilitation. Lower FIM® scores at admission to an IRF have been associated with an increased risk of being discharged to a non-community setting; in particular, several studies have found patients with a lower admission FIM® total had an increased odds of readmission to an acute care hospital [3,4]. Additionally, Tan et al. [4] found that admission FIM® total was a positive predictor of functional gain (score at discharge from score at admission) and length of stay in an IRF [5,6], [7,8]. Specific references included below.

The mobility measure was examined as a stand-alone measure, independent of the FIM instrument. The mobility measure items at admission, at discharge and the measure change score (difference in total score from admission to discharge) was compared to the total FIM Instrument scores at admission, discharge and change in FIM from admission to discharge score. The FIM Instrument has 18 items, 13 motor items and 5 cognitive items. Since the mobility measure does not include any cognitive items, only the 13 motor items of the FIM Instrument were used in the analysis.

The correlations between the mobility measure and the FIM motor items total were statistically significant ($p<.001$). The correlation between the admission mobility measure and the admission FIM motor total was .82, between the discharge mobility measure and the discharge FIM motor total was .93, and between the total change in mobility score and the total change in FIM motor score was .87.

The mobility measure, independently, was assessed to determine the predictive ability of the measure on patient outcomes. Predictive ability is of great importance in health care as it can be used to determine the relative influence, effect or contribution of a variable (such as level of function) upon another variable (like discharge to home or improvement in function) in order to detect which predictors have the strongest influence on outcomes. Predictive validity can be assessed using regression modeling. A R^2 value can be calculated which is interpreted as the proportion of variance accounted for, in essence, it explains how much of the variance in the dependent variable/outcome of interest (such as improvement in function), is accounted for by the independent variable/predictor (such as admission mobility score). The R^2 value is a number between 0 and 1 whereby a higher value indicate higher predictive validity. Regression modeling was performed to determine if the mobility measure is predictive of patient outcomes such as: change in function (total change in functional status from admission to discharge), and likelihood of discharge to the community setting from inpatient facility, and total length of stay (LOS) in inpatient facility. Linear regression was used to determine functional change, whereas the change in mobility score was the independent variable, the R^2 value (proportion of change accounted for) and the Pearson correlation coefficient were examined. For discharge to community setting, logistic regression was used, admission mobility total was the independent variable and the dependent variable was dichotomized as discharge to the community (yes or no). The adjusted R^2 value was examined to determine the proportion of variance accounted for by the measure. The regression models included all 4 items of the mobility measure.

The mobility measure was a significant predictor of patient discharge to the community, chi-square=46078.9, (df=4), $p<.001$, $R2 = .14$. The mobility measure was a significant predictor of patient LOS, adjusted $R2 = .15$, $p<.001$. The mobility measure was a significant predictor of patient functional change from admission to discharge, adjusted $R2 = .27$, $p<.001$. All 4 items of the mobility measure were retained in each of the regression models and were statistically significant ($p<.001$) in the models.

1. Medicare program; prospective payment system for inpatient rehabilitation facilities. Final rule. *Federal register.* Aug 7 2001;66(152):41315-41430.
2. Medicare program; inpatient rehabilitation facility prospective payment system for federal fiscal year 2012; changes in size and square footage of inpatient rehabilitation units and inpatient psychiatric units. Final rule. *Federal register.* Aug 5 2011;76(151):47836-47915.

3. Chung DM, Niewczyk P, Divita M, Markello S, Granger C. Predictors of discharge to acute care after inpatient rehabilitation in severely affected stroke patients. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* May 2012;91(5):387-392.

4. Tan WH, Goldstein R, Gerrard P, et al. Outcomes and predictors in burn rehabilitation. *Journal of burn care & research : official publication of the American Burn Association.* Jan-Feb 2012;33(1):110-117.

5. Inouye M, Hashimoto H, Mio T, Sumino K. Influence of admission functional status on functional change after stroke rehabilitation. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Feb 2001;80(2):121-125; quiz 126, 146.

6. Leung AW, Cheng SK, Mak AK, Leung KK, Li LS, Lee TM. Functional gain in hemorrhagic stroke patients is predicted by functional level and cognitive abilities measured at hospital admission. *NeuroRehabilitation.* 2010;27(4):351-358.

7. Franchignoni F, Tesio L, Martino MT, Benevolo E, Castagna M. Length of stay of stroke rehabilitation inpatients: prediction through the functional independence measure. *Annali dell'Istituto superiore di sanita.* 1998;34(4):463-467.

8. McClure JA, Salter K, Meyer M, Foley N, Kruger H, Teasell R. Predicting length of stay in patients admitted to stroke rehabilitation with high levels of functional independence. *Disability and rehabilitation.* 2011;33(23-24):2356-2361.

1a.3. SYSTEMATIC REVIEW(S) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	

Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 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.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

The current mandated quality measures for inpatient rehabilitation facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of inpatient rehabilitation is to increase patient function to return the patient to home/previous residence within the community. Yet the current measures don't adequately capture function or functional improvement. The current quality indicator measures address facility level process, which, has been argued, is not applicable to the inpatient medical rehabilitation setting as the overall prevalence of these events are very low (less than 2% of patients affected per year) and often times, the presence of the quality indicator occurred in the acute care setting or prior to admission to acute or post-acute care (for instance, CAUTIs and incidence of new or worsened pressure ulcers).

The mobility measure is constructed by utilizing items from the FIM® instrument, which is presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the FIM® instrument. The FIM® instrument is already used in inpatient

rehabilitation as it is embedded in the IRF-PAI, which is required to be completed for payment reimbursement by CMS. Each of the four items that comprise the mobility measure are presently collected in the IRF-PAI to capture patient functional status. Utilizing the change in mobility function measure as a quality indicator would not create any additional costs to IRFs, since IRFs are already transmitting the current IRF-PAI data to CMS for payment purposes. The change in mobility measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and is predictive of patient change in mobility function outcomes and likelihood of patient discharge from inpatient rehabilitation to home/the community. It is imperative that any quality indicators used in the post acute care setting take into account the overriding goal of medical rehabilitation, which is to restore and improve patient function and increase functional independence among individuals thus allowing the patient the ability to return to a community setting upon discharge from an inpatient facility.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Performance score results are detailed in the measure testing attachment.

Mean change in mobility scores at the facility level were computed and mobility change scores were grouped by quartile to determine if facilities can be 'ranked' in terms of patient outcomes (average change in mobility function from admission to discharge). There were 10 facilities in the 1st quartile (25th%) which includes mean mobility change scores less than 6.0, 538 facilities were in the 2nd quartile which includes mean mobility change scores of 6.0-9.9 (25th through 50th%), 197 facilities were in the 3rd quartile which includes mean mobility change scores of 10.0-13.0 (50th through 75th%) and 5 facilities were in the upper quartile (over 75th%) which includes mean mobility change scores greater than 13.0. An ANOVA was conducted using the quartiles as constructed above to determine if a statistically significant difference existed between the mobility change scores by quartile. The means and standard deviations are displayed below. There were statistically significant differences between the mean mobility change scores by quartile grouping, F=1073248.39 (df=3), p=.000. The Eta2 = .87. The Eta2 is the effect size; it is considered the most important outcome of empirical research because the effect size captures the practical significance of the research results. Eta2 is interpreted as the proportion of variance accounted for in the dependent variable (mean self-care change) that is associated with the membership of different groups in the independent variable (quartile) and the value is interpreted similar to a correlation coefficient where as a value of .2 is considered a small effect, .5 a moderate effect and .8 is a large, strong effect.

Mean Change in Mobility by Quartile

Quartile	Mean	N	Std. Deviation
25th%	2.7891	135779	2.58665
50th%	8.6365	140216	1.11181
75th%	11.4923	76065	.49994
over 75th%	15.2018	136882	2.02662
Total	9.2950	488942	5.07925

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.

Performance score results are detailed in the measure testing attachment and described above in 1b.2.

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 maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Results of disparities analysis is detailed in the measure testing attachment.

There were no differences in mean change in mobility score (change in total mobility score from admission to discharge) by race (eta2 <.001 for all race/ethnic categories), sex (eta2 <.001) or marital status (eta2 <.01).

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

Results of disparities analysis is detailed in the measure testing attachment and described above in 1b.4.

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 sub criteria 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):*

De.6. Non-Condition Specific*(check all the areas that apply):*

Care Coordination, Health and Functional Status : Change

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions

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.)*

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_Submission_Mobility-635533914241373843.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Attachment **Attachment:** IRFPAI_V20_2018-636903595864779736.pdf

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Clinician

S.3.1. <u>For maintenance of endorsement:</u> Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

No

S.3.2. <u>For maintenance of endorsement,</u> please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

No changes to the measure specifications since last endorsement.

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) DO NOT include the rationale for the measure.*

<u>IF an OUTCOME MEASURE</u>, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Average change in rasch derived mobility function score from admission to discharge at the facility level. Includes the following items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. Average is calculated as: (sum of change at the patient level/total number of patients). Patient less than 18 years of age at admission to the facility or patients who died within the facility are excluded.

S.5. 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, time period for data collection, 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)*

<u>IF an OUTCOME MEASURE,</u> describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

For Inpatient Rehabilitation Facilities (IRFs) data collection is presently required for payment reimbursement by the Centers for Medicare and Medicaid Services (CMS) using the mandated Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-PAI). Embedded in the IRF-PAI is the FIM® Instrument. The FIM® Instrument is a criterion referenced tool with 18 items that measures patient physical and cognitive function, need for helper assistance, burden of care/level of dependence. Each item is rated on a scale of 1 (most dependent) to 7 (completely independent). For the purposes of this measure, a subset of 4 FIM® items has been tested and validated as the Change in Mobility measure; the items are: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. Rasch analysis was performed on the items and the difference in the rasch derived values (defined in S.2b) from admission to discharge reflect the change at the patient level. The numerator of the measure is the average change in mobility score at the facility level.

While the IRF-PAI is specific to inpatient rehabilitation facilities, the change in mobility measure can be used in all post-acute care venues. The FIM® instrument is routinely used for patient functional assessment in all venues of care and has been tested and validated for use in IRFs, skilled nursing facilities (SNFs) and long term acute care facilities (LTAC) (www.udsmr.org), therefore this measure is not specific for inpatient medical rehabilitation use only.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

Facility adjusted expected change in rasch derived mobility values, adjusted at the Case Mix Group (CMG) level.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

To calculate the facility adjusted expected change in rasch derived mobility values, indirect standardization was used, which weights national CMG-specific values by facility-specific CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The case-mix group (CMG) classification system groups similarly impaired patients based on functional status at admission, in essence, patient severity. Patients within the same CMG are expected to have similar resource utilization needs and similar functional outcomes. There are three steps to classifying a patient into a CMG at admission:

1. Identify the patient's impairment group code (IGC).
2. Calculate the patient's weighted motor index score, calculated from 12 of the 13 motor FIM® items.
3. Calculate the cognitive FIM® rating and the patient´s age at admission. (This step is not required for all CMGs.)

See file uploaded in S.2b for calculations or ´CMG Version 3.00 [ZIP, 9.02mb]´ at the following link for more details:

https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/InpatientRehabFacPPS/CMG.html

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

National values used in the CMG adjustment procedure will not include cases who died in the IRF or patients less than 18 years of age at admission. Cases who died during rehabilitation are not typical patients and are routinely omitted from reports and published research on rehabilitation outcomes. Further details and references related to the exclusion criteria can be found in the Measure Testing form.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.)*

Patient date of birth (DOB), date of admission and discharge setting variables are collected in the IRF-PAI. Age can be calculated from DOB and admission date. The variable discharge setting includes a category for ´died´ which is indicated as a code of ´11´. Patient date of birth, admission date and discharge setting are also documented in SNFs and LTAC facilities.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.)*

While the measure can be stratified by specific impairment type (IGC), the CMG adjustment procedure allows for the measure to be complete, accurate, and valid for all patients within the facility, excluding died cases and ages less than 18.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Stratification by risk category/subgroup

If other:

S.12. Type of score:

Ratio

If other:

S.13. 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 = Higher score

S.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*

1. Target population: patients receiving care at an inpatient medical rehabilitation facility, a skilled nursing facility, or a long term acute care facility.
2. Exclusions: Age less than 18 years and patients who died during the episode of care.
3. Cases meeting target process: All remaining cases.
4. Outcome: Ratio of facility level average mobility change (rasch derived values) to facility CMG adjusted expected mobility change.
5. Risk adjustment: CMG adjustment using indirect standardization of the proportion of cases at the facility by CMG, and CMG specific national average of rasch derived value of mobility change.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

Measure is clinician assessed, proxy responses are not allowed.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

Measure is clinician assessed, not patient reported. All items are to be assessed on all patients aged 18 or older at admission and at discharge. All items are applicable and are required to be completed (items do not include a ´N/A´ or Missing category).

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Instrument-Based Data, Other

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

The items included in the Functional Change: Change in Motor Score measure are included in the FIM Instrument, which is embedded in the CMS IRF-PAI. The instrument is attached and can be accessed using the following link: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/Downloads/Final-IRF-PAI-Version-20-Effective-October-1-2018.pdf

Information related to assessment rules can be found under ´IRF-PAI Training Manual effective October 1, 2014 [ZIP, 2MB]´ using the following link: https://www.cms.gov/medicare/medicare-fee-for-service-payment/inpatientrehabfacpps/irfpai.html

S.19. 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.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility, Other

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital, Post-Acute Care

If other:

S.22. <u>COMPOSITE Performance Measure</u> - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

For inpatient rehabilitation facilities, CMGs were used to create the adjusted expectation. CMGs are comprised of: impairment group (IGC), functional status at admission based on 12 of the FIM items, and patient age at admission (for some CMGs). The FIM® instrument is divided into motor and cognitive items for CMG purposes. Twelve of the 13 motor items are used to calculate a weighted motor index. CMS created this weighting methodology as a way of accounting for the effect of each FIM® motor item on the cost of providing care to a patient in an IRF. The patient's weighted admission FIM® motor rating is the sum of the weighted admission ratings for the 12 FIM® motor items. The following weights are used for each item:

- Eating: 0.6
- Grooming: 0.2
- Bathing: 0.9
- Dressing – Upper Body: 0.2
- Dressing – Lower Body: 1.4
- Toileting: 1.2
- Bladder Management: 0.5
- Bowel Management: 0.2
- Transfers: Bed, Chair, Wheelchair: 2.2
- Transfers: Toilet: 1.4
- Locomotion: Walk, Wheelchair: 1.6
- Locomotion: Stairs: 1.6

CMS chose not to include the FIM item ´Transfers: Tub, Shower´ in the weighted motor score because analysis performed by the RAND Corporation for CMS found that this particular motor item did not contribute to the prediction of patient resource utilization as the other 12 FIM items did. When calculating the weighted admission FIM® motor rating, a score of 0 for ´Transfers: Toilet´ is converted to a score of 2; a score of 0 for any other item is converted to a score of 1.

While no such functional based grouping exists for LTAC facilities or for SNFs, this same process can be utilized in these other venues to group similarly functioning patients to allow for the adjusted comparison.

2. Validity – See attached Measure Testing Submission Form

NQF2321_Fall2018_testing_attachment_v7.1_Final.docx

2.1 <u>For maintenance of endorsement</u>

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.2 <u>For maintenance of endorsement</u>

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Yes - Updated information is included

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number: 2321
Measure Title: Functional Change: Change in Mobility Score
Date of Submission: 9/7/2019

Type of Measure:

☒ **Outcome (*including PRO-PM*)**	☐ **Composite – *STOP – use composite testing form***
☐ Intermediate Clinical Outcome	☐ Cost/resource
☐ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	

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: (*must be consistent with data sources entered in S.17*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ claims	☐ claims
☒ registry	☒ registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other:

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*).

Patient data from the FIM® instrument collected from inpatient rehabilitation facilities, long term acute care facilities, and skilled nursing facilities subscribing to the Uniform Data System for Medical Rehabilitation (UDSMR). The UDSMR maintains the largest non-governmental database for medical rehabilitation outcomes, whereby ~75% of all US inpatient rehabilitation facilities submit patient level data to include in facility level and national benchmarking reports. UDSMR is a not-for-profit organization affiliated with the University at Buffalo, located in Amherst, New York.

1.3. What are the dates of the data used in testing?

Patients discharged between 10/1/2016 to 9/30/2017 were included in the updated testing.

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: (*must be consistent with levels entered in item S.20*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: patient level change in function	☒ other: patient level/aggregate

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*)

All patients discharged from inpatient rehabilitation facilities (IRFs) in the U.S. between 10/1/2016 to 9/30/2017 (N=488,942, missing=0) were included in the updated aggregate testing. There were 855 facilities included, of which 76% were units within an acute care hospital and 24% were free-standing facilities. Every state in the U.S. was represented.

For the facility level analysis, a random sample of 30 facilities from the 855 total included facilities, were selected. A random sample was necessary as it would not be feasible to perform the analysis on all 855 facilities. The random sample included 7 freestanding IRFs and 23 units. Selection criteria for the random sampling were as follows: facilities must have had at least 100 cases discharged in the time period of reference and each facility contained complete patient records, meaning all items at admission and discharge were completed for each patient (no missing data). For the analysis, each of the 30 facilities were randomly split into two datasets, and the rasch-converted average change scores at the facility level were calculated, results were compared across the facilities (ICC).

To ensure the facilities selected in the random sample were representative of an average IRF, facilities with fewer than 100 patients discharged per year were excluded from selection, as this is not typical for the large majority of IRFs throughout the country and these facilities are outliers and likely differ in vast ways from the majority of IRFs. The number of facilities in the database that had less than 100 patients discharge in the one year time frame was 36, which is ~4% of the total number of facilities. Of the 36 facilities that were excluded from selection in the random sample, the average number of patients discharged per year was 74, with a median of 81, the range was 11 to 99. In contrast, of the 819 facilities eligible for selection in the random sample, the average number of patients discharged per year was 594, with a median of 413.

1.6. How many and which patients 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*)

All patients discharged from 855 facilities in the U.S. between 10/1/2016 to 9/30/2017 (N=488,942, missing=0) were included in the updated aggregate testing. Patient admission and discharge data were used. All patients age 18 and over were included, the mean age of the total sample was 69.1 years (S.D.= 15.5), 51% were female and 49% were male. All race/ethnicities were included and the distribution was as follows: 76.5% Caucasian (n=374,527), 12% African American (n=59,197), 6% Hispanic/Latino (n=28,321), 2% Asian (n=9,420), .5% Hawaiian/Pacific Islander (n=3,020), .5% American Indian/Alaskan (n=2,471), and 2.5% other race/ethnicity not specified (n=11,986). All payment sources were included and the distribution was as follows: 72.7% Medicare (n=355,424), 14.7% commercial health insurance (n=71,980), 6.5% Medicaid (n=31,717), 2.8% other payment source (Workers Compensation, no-fault auto, employer) (n=13,686), 1.7% unknown/payment source not specified (n=8,344), 1% un-reimbursed/no-pay (n=4,746), .3% Veterans benefits (n=1,545), .3% self-pay/private pay (n=1,500). All impairments/conditions were included; the distribution of sample impairment/conditions displayed in Table 1 below.

Table 1: Distribution of Sample by Impairment Type

Impairment Type	Frequency	Percent
Stroke	115607	23.6
Brain Dysfunction	55943	11.4
Neurologic Conditions	67436	13.8
Spinal Cord Dysfunction	28527	5.8
Amputation	15202	3.1
Arthritis	1877	.4
Pain Syndromes	1444	.3
Orthopedic Conditions	107219	21.9
Cardiac Disorders	22654	4.6
Pulmonary Disorders	7638	1.6
Burns	646	.1
Congenital Deformities	163	.0
Other Disabling Impairments	4421	.9
Major Multiple Trauma	14915	3.1
Debility	42571	8.7
Medically Complex Conditions	2679	.5
Total	488942	100.0

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.

N/A

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

The social risk variables available in the dataset include: race, sex and marital status.

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*)

SPSS version 22 was used to compute Cronbach's alpha to determine the internal consistency of the measure and to perform inter-item correlations.

Rasch analysis allows for the conversion of ordinal-level data into interval-level data. Ordinal measures do not inherently act as interval measures, where the difference between one score is equidistance compared to the difference between another two scores, i.e. the difference between a 15 and a 16 in our measure may not reflect the same difference between a 56 and a 57, in terms of difficulty. If the data fit the Rasch model, a result of the analysis is the conversion of the raw ordinal scores to a Rasch derived interval score. This allows for a more precise estimation of differences in functional status both between patients and across facilities. Rasch analysis was conducted to test the psychometric properties of the 4 items within the mobility measure and to determine the measure reliability at both the person and item level. Rasch analysis was also used to determine the fit of each item within the measure (4 items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) through infit and outfit statistics and item specific correlations. Winsteps 3.73 was used for the analysis.

To assess the measure reliability across facilities, an Intraclass Correlation Coefficient (ICC) using the split-half method was computed. The ICC analyses were previously suggested by the NQF PFCC committee staff. For the facility level analysis, a random sample of 30 facilities from the 855 included facilities, were selected. A random sample was necessary as it would not be feasible to perform the analysis on all 855 facilities. The random sample included 7 freestanding IRFs and 23 units. Selection criteria for the random sampling were as follows: facilities must have had at least 100 cases discharged in the time period of reference, each facility contained complete patient records, meaning all items at admission and discharge were completed for each patient (no missing data). For the analysis, each of the 30 facilities were randomly split into two datasets, and the rasch-derived average change score at the facility level was calculated, results were compared across the facilities (ICC).

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*)

The Cronbach Alpha reliability = .78, indicating a reliable measure (N=488,942, missing=0, number of items=4). Inter-item correlations ranged from .76 (transfer bed/chair and transfer toilet) to .37 (transfer toilet and walking), all items were significantly correlated (p <.001). Results of the Rasch analysis are as follows: The person-reliability correlation was 0.89. The infit and outfit statistics were acceptable for each of the 4 items (less than 2.0).

An intra-class correlation coefficient (ICC) using the split-half method was used to assess the score level reliability across facilities. The ICC was 0.951, p <.001. This high ICC demonstrates that there is very high consistency across facilities for the mobility measure. Rasch-converted average range in scores for the

measure by facility was 17.1 to 35.6. The average range in the mobility measure by facility is displayed in Table 2 below from lowest to highest value.

Table 2: Average Rasch-derived Mobility Scores by Facility

Facility	Score
Facility 28	17.14341
Facility 3	17.35923
Facility 8	18.15241
Facility 25	20.1361
Facility 23	20.68806
Facility 19	22.22116
Facility 14	22.59312
Facility 5	22.60908
Facility 17	23.21116
Facility 29	23.69272
Facility 13	23.70984
Facility 2	24.40056
Facility 4	24.47282
Facility 7	24.48544
Facility 1	24.86996
Facility 10	24.93485
Facility 18	24.95497
Facility 11	25.1936
Facility 15	25.99543
Facility 16	26.07491
Facility 20	27.5153
Facility 30	28.25309
Facility 22	28.92113
Facility 12	29.29216
Facility 6	29.36617
Facility 9	29.62186
Facility 21	31.14927
Facility 24	31.27187
Facility 26	34.24575
Facility 27	35.62555

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?*)

The results of the reliability analysis for the mobility measure were statistically significant; the Cronbach's alpha indicated high internal consistency, thus a stable measure. Inter item correlations were all statistically significant. The mobility measure is reliable and internally stable.

The facility-level intra-class correlation coefficient (ICC) was statistically significant demonstrating consistency among facilities in terms of ratings and outcome scores for the measure. Clearly there are differences in patient outcomes between facilities, as illustrated in the table above, whereby the average patient mobility

measure score (higher is better in terms of average patient function and facility performance) for Facility 28 is 17.1 vs 24.9 for Facility 1 vs 35.6 for Facility 27.

2b1. VALIDITY TESTING

2b1.1. 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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. 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)*

Construct validity is defined as the degree to which a test or instrument actually measures what it is intended the measure, the 'construct' of interest, for the current purpose, construct validity was assessed using SPSS to determine how well the mobility measure is able to capture the functional ability of a person to ambulate or move around the environment independently. Factor analysis using principal component analysis was used.

Predictive validity refers to the extent to which a score on an assessment predicts a future event, occurrence or performance. Predictive validity is of great importance as it can be used to determine the relative influence, effect or contribution of a variable (such as level of function) upon another variable (like discharge to home or improvement in function) in order to detect which predictors have the strongest influence on outcomes. Predictive validity can be assessed using regression modeling. A R^2 value can be calculated which is interpreted as the proportion of variance accounted for, in essence, it explains how much of the variance in the dependent variable/outcome of interest (such as improvement in function), is accounted for by the independent variable/predictor (such as admission mobility score). The R^2 value is a number between 0 and 1 whereby a higher value indicates higher predictive validity. Regression modeling was performed to determine if the mobility measure is predictive of patient outcomes such as: change in function (total change in functional status from admission to discharge), and likelihood of discharge to the community setting from inpatient facility, and total length of stay (LOS) in inpatient facility. Linear regression was used to determine functional change, whereas the change in mobility score was the independent variable, the R^2 value (proportion of change accounted for) and the Pearson correlation coefficient were examined. For discharge to community setting, logistic regression was used, admission mobility total was the independent variable and the dependent variable was dichotomized as discharge to the community (yes or no). The adjusted R^2 value was examined to determine the proportion of variance accounted for by the measure.

Criterion-referenced validity was assessed by comparing the mobility measure at admission, at discharge and the measure change score (difference in total score from admission to discharge) to the total FIM motor items scores (13 items) at admission, discharge and change in total motor FIM score from admission to discharge. The FIM Instrument is embedded in the IRF-PAI tool, which is used by IRFs throughout the country for reimbursement by CMS (mandated since 2002) and for patient level and facility level outcomes of care reporting, see udsmr.org for more information and for a list of references related to the FIM Instrument. The FIM Instrument has 18 items, 13 motor items and 5 cognitive items. Since the mobility measure does not include any cognitive items, only the 13 motor items of the FIM Instrument were used in the analysis.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Construct Validity

Factor analysis using principal component analysis resulted in 1 component identified in the mobility measure, cumulatively accounting for 61.1% of the total explained variance**.** Component 1 included items: transfer bed/chair (.86), transfer toilet (.84), walking (.69), and stairs (.73), eigenvalue=2.44.

Predictive Validity

Regression models were used to determine the predictive ability of the mobility measure items on patient outcomes. Specific patient outcomes included: patient discharge from inpatient facility to a community setting/home, total length of stay (LOS) in inpatient facility, and patient functional change from admission to inpatient facility to discharge. The regression models included all 4 items of the mobility measure. The mobility measure was a significant predictor of patient discharge to the community, chi-square=46078.9, (df=4), p<.001, R2 =.14. The mobility measure was a significant predictor of patient LOS, adjusted R2 =.15, p<.001. The mobility measure was a significant predictor of patient functional change from admission to discharge, adjusted R2= .27, p<.001. All 4 items of the mobility measure were retained in each of the regression models were and statistically significant (p<.001) in the models.

Criterion-referenced Validity

The correlations between the mobility measure and the FIM motor items total were statistically significant (p<.001). The correlation between the admission mobility measure and the admission FIM motor total was .82, between the discharge mobility measure and the discharge FIM motor total was .93, and between the total change in mobility score and the total change in FIM motor score was .87.

2b1.4. 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?*)

The results show the mobility measure is valid; the measure demonstrated construct, discriminant and predictive validity in all of the analyses. The r-square values in each of the regression models were moderate to strong, meaning the mobility measure was able to account for a significant percent of variance explained in the dependent variables. Overall, results of the principal components factor analysis were the mobility items cumulatively accounted for 61.1% of the total explained variance overall, which is strong and robust considering the very small number (4 items) of total items in the measure.

The results of the criterion-referenced validity testing indicated a very strong correlation between the mobility measure and the FIM motor total (13 items) at admission, discharge and the total change from admission to discharge. The very strong correlations with the FIM Instrument, the 'gold standard' measure for patient function, is evidence that the mobility measure, at just 4 items, is a predictive and robust measure of patient function and outcomes.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b4

2b2.1. 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*)

Patients that had expired in the inpatient rehabilitation facility (an unanticipated, very low frequency outcome) and patients under age 18 years were excluded from the analyses; both criteria are consistent exclusions in published literature examining rehabilitation outcomes.[3-8]

2b2.2. 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*)

Frequency of occurrence was less than 5% of total sample for each exclusion criteria (<1% died during inpatient stay and 4% of patients were under 18 years of age), findings are consistent with other published studies examining outcomes of inpatient medical rehabilitation using UDSMR data from earlier years. [5-7]

2b2.3. 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*)

Criteria for patient admission to inpatient rehabilitation includes patient ability to participate in three hours of intensive rehabilitation therapy per day every day (3 hrs per day at a minimum of 5 days per week) and a high likelihood for the patient to be discharged from the inpatient facility to their home or a community-based living setting (retirement community, assisted living, family member's home)[9]. Considering the aforementioned admission criteria, the 'typical' patient treated in inpatient rehabilitation tends to be medically stable and has functional limitations but shows some level of physical conditioning or functional abilities that indicate the individual would be able to withstand the three hours a day rigorous therapy requirement. Therefore, patients that are at increased risk for death in the near future (medically unstable/in critical condition, those requiring intensive 24 hour medical care, those with severe cognitive deficits (late stage dementia/Alzheimer's disease), patients with severe debility/highly deconditioned) do not tend to receive care in an inpatient rehabilitation facility and are more likely to be admitted to a skilled nursing facility, subacute facility, long-term acute care and/or Hospice care, so the small number of patients that do expire during an inpatient stay are extreme outliers and not representative of the larger population of patients that receive care at an inpatient rehabilitation facility.

Persons younger than age 18 requiring care at an inpatient rehabilitation facility may be treated at an adult facility or may receive care at one of many specialized pediatric facilities located throughout the country. Specialized pediatric inpatient rehabilitation facilities often provide a number of additional patient services such as educational coordination (in-hospital education/tutoring may be provided for school-age children) and developmental-related therapeutic services may be provided. Considering there may be a large difference in minors treated at a pediatric facility compared to those treated at a non-pediatric facility, and data are not available to UDSMR from the pediatric facilities for comparison, any results from persons under 18 years in the present dataset are not generalizable to the larger pediatric population and may be biased with a number of confounding variables. Thus, to control for possible bias, persons under 18 years of age were excluded from the present analyses.

2b3. 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**.**

2b3.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 1 risk factor**

☒ **Stratification by 1 risk categories**

☐ **Other,**

2b3.1.1 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.

The Case Mix Group (CMG) specifications, definitions, codes and algorithm can be accessed on the Centers for Medicare and Medicaid Services (CMS) website under the CMG version 3.00.[10] CMG version 3.00 can be accessed using the following link: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/InpatientRehabFacPPS/CMG.html.

2b3.2. 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

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

Outcomes, in terms of change in function, would lack meaning if *all* patients were aggregated, considering the many different conditions that patients are admitted to an inpatient rehabilitation facility for treatment. There are vast differences in patient conditions (stroke vs spinal cord injury vs knee replacement) in addition to a range of severity of within each condition (ex. for spinal cord injury there is a large difference in functional impairment between a patient with quadriplegia vs a patient with central cord syndrome), therefore, when examining functional outcomes of inpatient care it is imperative to control for these difference both between patient impairment types/conditions and to control for the severity within a given impairment type/condition. Stratification for patient impairment type/condition and risk adjusting data by CMG has been used extensively in prior, published research on patient functional outcomes of inpatient rehabilitation.[11, 12] CMG adjustment is a standard and expected procedure[10, 12].

Data was first stratified by impairment type, which is the specific, primary condition/reason a person is admitted to inpatient rehabilitation. The impairment types are listed in Table 1, previously displayed.

Next, the data was adjusted by Case Mix Group (CMG) through indirect standardization.

CMG is a proxy for severity of condition, just as two conditions are very different in terms of physical, psychological, physiological, cognitive and quality of life impact, there can be a large difference in severity within the same condition, for instance, a stroke can be very mild where no limitations in physical or cognitive functioning occur or a stroke can be very severe, whereby if death does not occur the result may include major cognitive impairment, loss of ability to control facilities, loss of speech, inability to swallow, walk or dress self.

To calculate the facility's adjusted expected change in Rasch derived values, indirect standardization was used which weights national CMG-specific values by facility-specific CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The CMG classification system groups similarly impaired patients based on functional status at admission or patient severity[10]. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes. There are three steps to classifying a patient into a CMG at admission:

1. Identify the patient's impairment group code (IGC).
2. Calculate the patient's weighted motor index score, calculated from 12 of the 13 motor FIM® items.
3. Calculate the cognitive FIM® rating and the age at admission (this step is not required for all CMGs, see specifications on CMS website for more details). [10]

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

☒ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

Published literature was used to determine appropriate risk factors to adjust for in the dataset[1-12].

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk 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.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

There were no differences in mean change in mobility score (change in total mobility score from admission to discharge) by race (eta^2 <.001 for all race/ethnic categories), sex (eta^2 <.001) or marital status (eta^2 <.01). This is consistent with the findings of other published research.[1, 2]

2b3.5. 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*)

CMG adjustment is a standard procedure[4, 5]. Stratification for patient impairment type/condition and risk adjusting data by CMG has been used extensively in prior, published research on patient functional outcomes of inpatient rehabilitation[4-12].

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Due to constraints in space/size the full CMG analysis is not included in the present testing submission. However, to illustrate the risk adjustment by CMG, one impairment type was selected and presented below. Considering stroke was the most frequently occurring impairment in the dataset (23.6% of patients (n=115,607) had stroke as their admission condition), stroke was selected and the data was stratified by impairment and only cases with stroke were included in the CMG risk adjustment analysis. An ANOVA was conducted to determine if there were significant differences in mean mobility change by CMG within the stroke subset. CMG for stroke includes 10 categories, ranging from 101 to 110, whereby CMG 101 is the least severe and CMG 110 is the most severe. A statistically significant difference was found in mean mobility change by CMG, F=452.82 (df=9), p=.000. The mean and standard deviation in total mobility change (total mobility score from admission to discharge) by CMG for patients with stroke is displayed in Table 3 below.

Table 3: Mean Mobility Change Score by CMG for Patients with Stroke

CMG		Mobility Change
101.00	Mean	6.9804
	N	5457
	Std. Dev	3.24192
102.00	Mean	7.9247
	N	7427
	Std. Dev	3.39505
103.00	Mean	7.0223
	N	2166
	Std. Dev	3.69355
104.00	Mean	8.4099
	N	12900
	Std. Dev	3.76993
105.00	Mean	8.9847
	N	11049
	Std. Dev	4.05513
106.00	Mean	9.3334
	N	10414
	Std. Dev	4.30857
107.00	Mean	9.5344
	N	10072
	Std. Dev	4.57769
108.00	Mean	6.9550
	N	8441
	Std. Dev	5.31147
109.00	Mean	9.9574
	N	9037
	Std. Dev	4.79376
110.00	Mean	7.8708
	N	38644
	Std. Dev	5.79078
Total	Mean	8.3087
	N	115607
	Std. Dev	4.94022

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

N/A

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

N/A

2b3.9. Results of Risk Stratification Analysis:

An ANOVA was conducted to determine if there were significant differences in mean mobility change by impairment type. A statistically significant difference was found in mean mobility change by impairment type,

F=1021.40 (df=15), p=.000. The mean and standard deviation in total mobility change (total mobility score from admission to discharge) by impairment type is displayed in Table 4 below.

Table 4: Mean Change in Mobility by Impairment Type

Impairment Type	Mean	N	Std. Dev.
Stroke	8.3087	115607	4.94022
Brain Dysfunction	8.4708	55943	5.09892
Neurologic Conditions	9.4318	67436	5.21427
Spinal Cord Dysfunction	9.2613	28527	5.33507
Amputation	8.9893	15202	4.92840
Arthritis	10.0703	1877	5.01685
Pain Syndromes	9.7126	1444	4.92154
Orthopedic Conditions	10.6919	107219	4.78921
Cardiac Disorders	9.2988	22654	4.92318
Pulmonary Disorders	8.7867	7638	4.83835
Burns	10.4536	646	5.42519
Congenital Deformities	8.7532	163	5.26820
Other Disabling Impairments	9.3646	4421	4.93795
Major Multiple Trauma	10.8115	14915	5.09728
Debility	9.0127	42571	4.99517
Medically Complex Conditions	8.1385	2679	4.93761
Total	9.2950	488942	5.07925

2b3.10. 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*)

There are statistically significant differences in mean mobility change by impairment type and by CMG; the mobility measure is able to discriminate in functional ability both between different functional impairments and within the same type of functional impairment (such as stroke).

2b3.11. 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*)

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. 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)*

Mean change in mobility scores at the facility level were computed and mobility change scores were grouped by quartile to determine if facilities can be 'ranked' in terms of patient outcomes (average change in mobility function from admission to discharge).

2b4.2. 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)

There were 10 facilities in the 1st quartile (25th%) which includes mean mobility change scores less than 6.0, 538 facilities were in the 2nd quartile which includes mean mobility change scores of 6.0-9.9 (25th through 50th%), 197 facilities were in the 3rd quartile which includes mean mobility change scores of 10.0-13.0 (50th through 75th%) and 5 facilities were in the upper quartile (over 75th%) which includes mean mobility change scores greater than 13.0. An ANOVA was conducted using the quartiles as constructed above to determine if there is a statistically significant difference between the mobility change scores by quartile. The means and standard deviations are displayed in Table 5 below. There were statistically significant differences between the mean mobility change scores by quartile grouping, F=1073248.39 (df=3), p=.000. The $Eta^2 = .87$. The Eta^2 is the effect size; it is considered the most important outcome of empirical research because the effect size captures the practical significance of the research results[15]. Eta^2 is interpreted as the proportion of variance accounted for in the dependent variable (mean self-care change) that is associated with the membership of different groups in the independent variable (quartile) and the value is interpreted similar to a correlation coefficient where as a value of .2 is considered a small effect, .5 a moderate effect and .8 is a large, strong effect[15].

Table 5: Mean Change in Mobility by Quartile

Quartile	Mean	N	Std. Deviation
25th%	2.7891	135779	2.58665
50th%	8.6365	140216	1.11181
75th%	11.4923	76065	.49994
over 75th%/upper quartile	15.2018	136882	2.02662
Total	9.2950	488942	5.07925

2b4.3. 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?*)

Difference in average mobility change scores between facilities can be determined and rank ordered in terms of patient average change in mobility function from admission to discharge. From the above mentioned results, clearly 'top performing' facilities, in terms of patient change in function, can be identified, in addition to facilities that are at the lowest quartile. There were statistically significant differences in mean change scores by quartile and the standard deviations within the quartiles were small, indicating some variability within groups but small enough so the scores are fully contained within the quartile and do not extend into another category. The Eta^2 value is very strong, which is further evidence that the differences in mean scores are true differences and not a result of the very large sample size; a very large sample can often can lead to small, negligible differences detected as statistically significant but when examining the actual values, the differences are not clinically relevant or meaningful (for instance a difference in self-care change of6.2 and 6.4, may be statistically significant due to the very large sample size but both values are, in essence, a 6, so the difference is not clinically relevant or meaningful in any way).

2b5. 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 social risk 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 social risk 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.***

2b5.1. 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*)

2b5.2. 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*)

2b5.3. 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*)

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. 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*).

There were no missing data in the dataset. The items in the mobility measure were selected from the FIM instrument, which is embedded in the IRF-PAI instrument[13], used by inpatient rehabilitation facilities for prospective payment by CMS. The instructions for rating the specific items can be found in the IRF-PAI Training Manual[14], effective 10/1/2014, and can be accessed using the following link: https://www.cms.gov/medicare/medicare-fee-for-service-payment/inpatientrehabfacpps/irfpai.html. In brief, the item rating rules state that a code of 0 may be used for the items at admission if the activity did not occur. It is stated that use of this code should be rare and that a code of 0 translates to a 1 (most dependent) in facility level reports and aggregate patient data reports. A code of 0 may not be used for for any items at discharge. All items are to be rated at both admission and discharge and there are no options for 'not applicable' or 'do not apply' or 'missing data' codes.

2b6.2. 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*)

No missing data in dataset/analyses.

2b6.3. 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*)

No missing data in dataset/analyses.

References

1. Rabadi M, Rabadi F, Hallford G, Aston C. Does Race Influence Functional Outcomes in Patients with Acute Stroke Undergoing Inpatient Rehabilitation? *PM&R.* 2012; 91(5):375–386.

2. Graham J, Radice-Neumann D, Reistetter T, Hammond F, Dijkers M, Granger C. Influence of sex and age on inpatient rehabilitation outcomes among older adults with traumatic brain injury. *Arch Phys Med Rehabil*. 2010; 91(1):43-50.

3. Gerrard P, Goldstein R, Divita M, et al. Validity and Reliability of the FIM(R) Instrument in the Inpatient Burn Rehabilitation Population. *Arch Phys Med Rehabil.* 2013; 94(8):1521-1526.

4. Cary M, Merwin E, Oliver M, Williams I. Inpatient Rehabilitation Outcomes in a National Sample of Medicare Beneficiaries With Hip Fracture. *J Appl Gerontol*. 2014; 35(1):62-83.

5. Ramey L, Goldstein R, Zafonte R, Ryan C, Kazis L, Schneider J. Variation in 30-Day Readmission Rates Among Medically Complex Patients at Inpatient Rehabilitation Facilities and Contributing Factors. *JAMDA*. 2016; 17:730-736.

6. Mix J, Granger C, LaMonte M, Niewczyk P, DiVita M, Goldstein R, Yates J, Freudenheim J. Characterization of Cancer Patients in Inpatient Rehabilitation Facilities: A Retrospective Cohort Study. *Arch Phys Med Rehabil.* 2017; 98(5):971-980.

7. DiVita M, Granger C, Goldstein R, Niewczyk P, Freudenheim J. Mandated Quality of Care Metrics for Medicare Patients: Examining New or Worsened Pressure Ulcers and Rehabilitation Outcomes in United States Inpatient Rehabilitation Facilities. *Arch Phys Med Rehabil.* 2018; 99(8):1514-1524.

8. Bajorek A, Slocum C, Goldstein R, Mix J, Niewczyk P, Ryan C, Hendricks C, Zafonte R, Schneider J. Impact of Cognition on Burn Inpatient Rehabilitation Outcomes. *PM&R.* 2017; 9(1): 1-7.

9. Duncan, P, Horner, Reker et al. Adherence to Post Acute Rehabilitation Guidelines is Associated with Functional Recovery in Stroke. *Strok.,* 2002; 33:167-178.

10. Centers for Medicare and Medicaid Services (CMS). Inpatient Rehabilitation Facility PPS. IRF Grouper- Case Mix Group (CMG). CMG version 3.00. Accessible at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/InpatientRehabFacPPS/CMG.html

11. Stineman M, Shea J, Jette A, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Arch Phys Med Rehabil.* 1996; 77(11):1101-1108.

12. Stineman MG, Granger CV. A modular case-mix classification system for medical rehabilitation illustrated. *Health Care Financ Rev*. 1997; 19(1):87-103.

13. Centers for Medicare and Medicaid Services. Inpatient Rehabilitation Facility PPS. IRF-PAI. IRF-PAI version 1.4, effective October 1, 2016. Accessible at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/InpatientRehabFacPPS/Downloads/IRF-PAI-Version-1-4.pdf.

14. Centers for Medicare and Medicaid Services. Inpatient Rehabilitation Facility PPS. IRF-PAI. 2014 IRF-PAI Information, IRF-PAI Training Manual, effective October 1, 2014. Accessible at: https://www.cms.gov/medicare/medicare-fee-for-service-payment/inpatientrehabfacpps/irfpai.html.

15. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVA. *Front Psychol.* 2013; 4:863.

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.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

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*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic clinical data (e.g., clinical registry, nursing home MDS, home health OASIS)

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. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

Presently, no efforts to develop an eMeasure for the Change in Mobility Score Measure.

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. Please also complete and attach the NQF Feasibility Score Card.

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. Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Collection of the mobility items has occurred in IRFs for nearly thirty years via the FIM Instrument. UDSMR has data beginning in 1987 on the items within the change in mobility measure. Since the FIM instrument is required for payment via IRF-PAI, there are no missing data on any items within the measure. All patients treated in an IRF are administered the FIM instrument upon admission and at discharge, and therefore there is no additional time or cost associated with implementing this measure.

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)*.

Measure is publicly available and free of charge for use. Facility-level benchmark reporting is available through UDSMR via subscription, cost varies by facility type and size. National reporting could be available free of charge if CMS elects to provide.

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.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Regulatory and Accreditation Programs	Payment Program CMS IRF-PAI cms.gov Quality Improvement (external benchmarking to organizations) UDSMR IRF Benchmarking Reports udsmr.org Quality Improvement (Internal to the specific organization) udsmr.org UDSMR IRF Facility-level Reports

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

Completion of the IRF-PAI is required by all IRFs throughout the country for all patients in which the IRF requests payment reimbursement from the Inpatient Rehabilitation Facility Prospective Payment System through the Centers for Medicare and Medicaid Services.
FIM® outcomes are currently benchmarked for all UDSMR subscribing facilities, with internal and external benchmarking options. UDSMR has 875 current enrolled IRFs which is roughly 80% of all IRF in the U.S. In addition, there are SNFs and LTAC facilities that subscribe to UDSMR and utilize the FIM® instrument to track patient functional outcomes (SNF = 152 and LTAC = 7).

4a1.2. 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?*)
As described above in 4a1.1, completion of the IRF-PAI is required by all IRFs throughout the country for all patients in which the IRF requests payment reimbursement from the Inpatient Rehabilitation Facility Prospective Payment System through the Centers for Medicare and Medicaid Services.

4a1.3. 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.*)

As described above in 4a1.1, completion of the IRF-PAI is required by all IRFs throughout the country for all patients in which the IRF requests payment reimbursement from the Inpatient Rehabilitation Facility Prospective Payment System through the Centers for Medicare and Medicaid Services.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Facilities subscribing to UDSMR primarily do so because of the national benchmarking reports and services that UDSMR provides. Patient outcomes are currently benchmarked for all UDSMR subscribing facilities, with facility level and national benchmark reporting provided on a quarterly basis. UDSMR maintains the world's largest government-independent repository of rehabilitation outcomes and IRF-PAI data. The repository contains data from over 1,400 rehabilitation facilities worldwide, 875 of which are IRFs in the United States, that use UDSMR's outcomes reporting, credentialing, auditing, training, and consulting services. UDSMR works with subscribing facilities and healthcare providers to document and improve patient functional outcomes, facility-level quality processes, and delivery of care in a uniform, standardized way.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

Patient level outcomes are currently benchmarked for all UDSMR subscribing facilities, with facility level and national benchmark reporting provided on a quarterly basis. UDSMR works with subscribing facilities and healthcare providers to document and improve patient functional outcomes, facility-level quality processes, and delivery of care in a uniform, standardized way. UDSMR provides assistance to subscribing facilities on coding and assessment education and support. Training for facility providers on patient assessment and functional outcomes documentation of the functional measures. Custom reports, facility quality improvement information and report interpretation is provided by request at no charge.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Feedback from users was not solicited. The items in the measure are not new and have been in use for over two decades in post acute rehabilitation.

4a2.2.2. Summarize the feedback obtained from those being measured.

Feedback was not solicited. The measure is clinician assessed and collected routinely as part of clinical care. Patients are not questioned and do not provide any responses for the items within the measures. The items in the measure are not new and have been in use for over two decades in post acute rehabilitation.

4a2.2.3. Summarize the feedback obtained from other users

No feedback was received from other users.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

No modifications were made to the measure. No negative feedback was received.

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.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)

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.

Difference in average mobility change scores between facilities can be determined and rank ordered in terms of patient average change in mobility function from admission to discharge. Top performing facilities, in terms of patient change in function, can be identified, in addition to facilities that are at the lowest quartile. There were statistically significant differences in mean change scores by quartile and the standard deviations within the quartiles were small, indicating some variability within groups but small enough so the scores are fully contained within the quartile and do not extend into another category. The Eta2 value is very strong, which is further evidence that the differences in mean scores are true differences and not a result of the very large sample size; a very large sample can often can lead to small, negligible differences detected as statistically significant but when examining the actual values, the differences are not clinically relevant or meaningful.

4b2. 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).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

There were no unexpected or unanticipated findings.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

There are statistically significant differences in mean mobility change by impairment type and by CMG; the mobility measure is able to discriminate in patient functional ability both between different functional impairments and within the same type of functional impairment (such as stroke).

The results of the criterion-referenced validity testing indicated a very strong correlation between the mobility measure and the FIM motor total (13 items) at admission, discharge and the total change from admission to discharge. The very strong correlations with the FIM Instrument, the 'gold standard' measure for patient function, is evidence that the mobility measure, at just 4 items, is a predictive and robust measure of patient function and outcomes.

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.

No

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

5a. Harmonization of Related Measures

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 harmonized to the extent possible?

No

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

Measure 2321 includes 4 items rated on a 7 level scale, clinicians rate the patient´s lowest actual observed score over the past 24 hour period (if patient is independent with toileting while awake but needs assistance in the middle of the night the rating would be the lowest/middle of the night score for the item, all items are to be rated at admission and at discharge, there are no codes for missing/do not apply. Measure 2634 includes 15 mobility items, many of which measure the same construct such as ambulation, but capture the distance walked (10 feet, 50 feet) as separate items, opposed to measure 2321 whereby the distance is captured within the item as part of the rating scale. Measure 2634 uses a 6 level rating scale (1-6) and includes options for not assessing each item, thus allows for missing responses (ex. not applicable/ patient refused/ not attempted due to safety), the patient´s usual performance is used as the basis of the score where if a patient were independent in toileting during the day but needed assistance in the middle of the night the score would be independent as there would be more frequent independent episodes throughout the day opposed to a single instance over night. These measures use different rating scales and different assessment rules and when trying to determine a patient´s actual level of function, a 6 level scale is less sensitive than a 7 level scale as there is less ´room´ to demonstrate change over time captured in the 6 level rating scale. Additionally, if determining patient discharge setting, using ´patient usual performance´ may portray a higher level of function than truly exists for the patient, whereby if it is believed the patient is independent in certain items but does in fact need assistance at certain times of day or in some instances, and there are not provisions in place to provide the care, the patient is at risk for a fall or readmission to inpatient care if a caregiver or attendant is not with the patient to provide the assistance (such as in the example of toileting used previously). Furthermore, the inclusion of multiple ´missing´ options for each item to be allowed for use at admission and at discharge lends the possibility for data that is not able to be interpreted, if an item is not rated at admission because the patient refused but is rated at discharge, of what value is this information? It is unknown if the patient would have been rated the exact same at admission, thus no change actually occurred from admission to discharge, of if there were an improvement, it would not be captured, or if there was a decline in function, this too is unknown, so if an item is not applicable (or not safe for administration at admission) than it lends question as to why it is included in the measure at all and if it is applicable, allowing missing values adds to the clinical data collection burden without any benefit to the patient as any other values collected cannot be interpreted directly when an item was missing at another point in time (an admission rating but no discharge rating or vice versa). Predictive models at the measure level require complete data so even if one value is missing for one item the entire case is dropped from the analytical model.

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.)

Measure #2321 is similar to CMS Measure #2634, however Measure #2634 is only intended for Medicare patients (majority of which are age 65 or older) treated at an IRF whereas Measure #2321 is intended for all

patients age 18 and older receiving post acute care at an IRF, SNF or LTAC facility. Measure #2321 includes four mobility items, whereby Measure #2634 includes 15 mobility items, several of which are redundant and may add to patient and clinician assessment burden. Furthermore, several of the items are not feasible for patients in an inpatient setting, such as the following items: car transfer, walk on uneven surfaces, bend to pick up an object while standing, especially upon admission. This is acknowledged considering there are four missing codes for all of the mobility items (patient refused, not applicable, not attempted due to safety concerns and not attempted due to environmental limitations). Measure #2321 is applicable for all adult patients and is intended to be assessed for all adult patients at both admission and discharge. If an item is not applicable at admission, a change score cannot be computed and true assessment of patient and facility outcomes may be biased based on the missing data. Furthermore, true validation of the measure requires complete data for all items within the measure, otherwise cases with even just one item missing are eliminated from the statistical model. This may result in a large amount of missing data compared to the total number of cases assessed and the results of the analysis would be biased to include only complete cases with no missing data, these cases are likely VERY different (and much higher functioning, if a patient can walk on an uneven surface at admission to an IRF) than other patients where the given item(s) was not attempted at admission (more typical in terms of the type of patient admitted to an IRF).

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:** Appendix-636824836707822691-636845229814709570.pdf

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.

Co.2 Point of Contact: Paulette, Niewczyk, pniewczyk@udsmr.org, 716-817-7868-

Co.3 Measure Developer if different from Measure Steward: Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.

Co.4 Point of Contact: Paulette, Niewczyk, pniewczyk@udsmr.org, 716-817-7864-

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.

Margaret DiVita, PhD, UDSMR assisted with measure testing.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2014

Ad.3 Month and Year of most recent revision: 04, 2019

Ad.4 What is your frequency for review/update of this measure? Unknown, new measure

Ad.5 When is the next scheduled review/update for this measure? 04, 2019

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: