**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** Not applicable

**Measure Title**: Inpatient Rehabilitation Facility (IRF) Functional Outcome Measure: Change in Self-Care Score for Medical Rehabilitation Patients

**Date of Submission**: 11/3/2014

**Type of Measure:**

|  |  |
| --- | --- |
| Composite – ***STOP – use composite testing form*** | x Outcome (*including PRO-PM*) |
| ☐ Cost/resource | ☐ Process |
| ☐ Efficiency | ☐ Structure |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specifications** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| ☐ administrative claims | ☐ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| ☐ abstracted from electronic health record | ☐ abstracted from electronic health record |
| ☐ eMeasure (HQMF) implemented in EHRs | ☐ eMeasure (HQMF) implemented in EHRs |
| X other: CARE Tool used during the Post-Acute Care Payment Reform Demonstration – electronic submission. | X other: CARE Tool used during the Post-Acute Care Payment Reform Demonstration – electronic submission. |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

The datasets used for the testing of this quality measure are: 1) the CARE Tool data collected during the Post-Acute Care Payment Reform Demonstration and 2) Medicare claims data.

**Citation:**

Centers for Medicare & Medicaid Services. (2013, July). *CARE Item Set and B-CARE*. Retrieved from <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/CARE-Item-Set-and-B-CARE.html>

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

Data used in testing were collected during the years 2008 to 2010.

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all 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.26*)** | **Measure Tested at Level of:** |
| ☐ individual clinician | ☐ individual clinician |
| ☐ group/practice | ☐ group/practice |
| X hospital/facility/agency | X hospital/facility/agency |
| ☐ health plan | ☐ health plan |
| ☐ other: Click here to describe | ☐ other: Click here to describe |

**1.5. How many and which measured entities 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*)

Testing for this quality measure involved several types of item- and scale-level reliability and validity analyses. The IRF self-care function quality measure includes data for 7 items that are summed to create a scale score. Each provider collected CARE data for either 6 or 9 months (not 12 months, which is the measure time period), so we provide only limited facility-level analyses. The CARE self-care and mobility function items were tested in several types of providers: acute care hospitals, Skilled Nursing Facilities, Long-Term Care Hospitals, and Home Health Agencies in addition to Inpatient Rehabilitation Facilities (IRFs). We report reliability and validity results overall and by type of provider.

**Item Reliability Testing - Traditional Interrater Reliability:** The interrater reliability of the CARE function items were tested in a sample of 34 providers distributed across 11 geographic areas: Lakeland/Tampa, FL (3 providers); Lincoln/Omaha, NE (5 providers); Louisville, KY (4 providers); Chicago, IL (5 providers); Dallas, TX (6 providers); Wilmington, NC (2 providers); Columbia, MO (2 providers); Seattle, WA (2 providers); San Francisco, CA (3 providers); Boston, MA (1 provider); and Rochester, NY (1 provider). The type of providers included: 4 acute hospitals (66 paired assessments); 8 Home Health Agencies (102 paired assessments); 7 Inpatient Rehabilitation Facilities (119 paired assessments); 2 Long-Term Care Hospitals (49 paired assessments); and 6 Skilled Nursing Facilities (121 paired assessments).

**Item Reliability Testing - Videotaped Standardized Patient Interrater Reliability:** The standardized patient reliability was tested in a sample of 28 providers (550 assessments), which included 3 acute hospitals (15 assessments [3%]); 9 Home Health Agencies (118 assessments [22%]); 8 Inpatient Rehabilitation Facilities (237 assessments [43%]); 3 Long-Term Care Hospitals (114 assessments [21%]); and 5 Skilled Nursing Facilities (66 assessments [12%]).

**Scale Reliability and Construct Validity Testing - Item Difficulty Ordering (Rasch Analysis):** Data from a sample of 38 IRFs were used for the Rasch analysis response option assessment. The sample was distributed across 16 geographic areas: Boston, MA; Chicago, IL; Dallas, TX; San Francisco/Sacramento, CA; Seattle, WA/Portland. OR; Lincoln/Omaha, NE; Lakeland/Tampa, FL; Louisville, KY; Wilmington, NC; Rochester, NY; New Hampshire/Maine; New York/New Jersey/Pennsylvania; Virginia; Maryland/Washington, DC; Ohio/Michigan; California.

**Scale Reliability and Validity Testing - Fit Assessment (Rasch Analysis) and Internal Consistency:** Data from a sample of 38 IRFs were used for the Rasch analysis fit assessment. The sample was distributed across 16 geographic areas: Boston, MA; Chicago, IL; Dallas, TX; San Francisco/Sacramento, CA; Seattle, WA/Portland. OR; Lincoln/Omaha, NE; Lakeland/Tampa, FL; Louisville, KY; Wilmington, NC; Rochester, NY; New Hampshire/Maine; New York/New Jersey/Pennsylvania; Virginia; Maryland/Washington, DC; Ohio/Michigan; California.

**Item Reliability and Validity Testing - Response Option Assessment (Rasch Analysis):** Data from a sample of 38 IRFs were used for the Rasch analysis response option assessment. The sample was distributed across 16 geographic areas: Boston, MA; Chicago, IL; Dallas, TX; San Francisco/Sacramento, CA; Seattle, WA/Portland. OR; Lincoln/Omaha, NE; Lakeland/Tampa, FL; Louisville, KY; Wilmington, NC; Rochester, NY; New Hampshire/Maine; New York/New Jersey/Pennsylvania; Virginia; Maryland/Washington, DC; Ohio/Michigan; California.

**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*)

**Item Reliability Testing – Traditional Interrater Reliability:** Providers (all types of providers) were asked to enroll a convenience sample of Medicare patients representing a range of clinical and functional complexity. The overall patient sample size for all but one self-care item was 450; it was 449 for Eating. After excluding the records of patients with “not attempted” codes and missing data, the effective sample sizes for the reliability testing were: Eating: 401; Oral hygiene: 414; Toileting hygiene: 416; Upper body dressing: 420; Lower body dressing: 413. Patients in the interrater reliability sample had a mean age of 77 and the majority of patients were female (59.6%) and white (89.9%). More than half of the sample (65.5%) was admitted directly from a short-stay acute hospital, and 14.7% were admitted directly from the community.

**Item Reliability Testing - Videotaped Standardized Patient Interrater Reliability:** For the video reliability testing, which was designed to examine clinician agreement across provider types, clinicians were asked to assess “standardized” patients presented through a videotape of a patient assessment. This ensured that the same information was presented to each clinician and allowed examination of differences in scoring effects among different clinicians examining the “same” patient. The patient “case studies” in each of the videos varied in terms of medical complexity, functional abilities, and cognitive impairments. The nine videos included patients classified as high, medium, or low ability/complexity for each of these three areas. Each facility or agency received three videos, one of which demonstrated one of the following elements: cognitive impairments, skin integrity problems, a wheelchair-dependent patient, and a variety of mid-level functional activities.

**Scale Reliability and Construct Validity Testing - Item Difficulty Ordering (Rasch Analysis):** Data for 5,129 IRF patients was analyzed for the fit assessment and internal consistency. More than half of the IRF patients were female (57.9%) and 57.6% were 75 years old or older. Most (87.7%) were white and 7.6% had Medicaid as a secondary payer. Most (92.1%) were admitted to the IRF directly from an acute care hospital.

**Scale Reliability and Validity Testing - Fit Assessment (Rasch Analysis) and Internal Consistency:** Data for 5,129 IRF patients was analyzed for the fit assessment and internal consistency. More than half of the IRF patients were female (57.9%) and 57.6% were 75 years old or older. Most (87.7%) were white and 7.6% had Medicaid as a secondary payer. Most (92.1%) were admitted to the IRF directly from an acute care hospital.

**Item Reliability and Validity Testing - Response Option Assessment (Rasch Analysis):** Data for 5,129 IRF patients was analyzed for the response option analysis (Rasch analysis). More than half of the IRF patients were female (57.9%) and 57.6% were 75 years old or older. Most (87.7%) were white and 7.6% had Medicaid as a secondary payer. Most (92.1%) were admitted to the IRF directly from an acute care hospital.

**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**.

All testing was conducted as part of the Post-Acute Care Payment Reform Demonstration; each type of testing involved a different set of providers. The traditional interrater reliability testing included a total of 34 providers, of which 7 were IRFs with 119 paired IRF assessments. A total of 9 videotaped “standardized” patients were used to test interrater reliability in 8 IRFs with 238 IRF assessments. The IRF sample for the Rasch analyses and internal consistency analyses was 5,129 patients from 38 IRFs.

**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*)  
X **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*)

**Item Reliability - Traditional Interrater Reliability and** **Standardized Patients Interrater Reliability:** Two types of interrater reliability testing were conducted to examine item (i.e., data element) reliability. For the traditional interrater reliability analyses, patients were assessed by two different clinicians at the same provider and the agreement of clinicians’ scoring was calculated. Providers were instructed to have pairs of clinicians complete both patient assessments at the same time. The second type of interrater reliability testing examined agreement of clinicians’ scores across different providers (i.e. different IRFs) and different types of providers. The videotaped “standardized” patient analyses involved clinicians in different settings rated items based on watching the “standardized” patients, and agreement of clinicians’ scoring was calculated.

**Scale Reliability Testing - Internal Consistency:** We used internal consistency to **p**rovide a general assessment of how well the items interrelate within a domain or subscale.

**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*)

**Item Reliability - Traditional Interrater Reliability:** RTI used two analytic approaches for assessing the interrater reliability of the CARE function items. For continuous items, RTI calculated Pearson correlation coefficients to show the extent of agreement between two clinicians on the same item. For categorical items, RTI calculated kappa statistics which indicate the level of agreement between raters using ordinal data, taking into account the role of chance agreement. The ranges commonly used to judge reliability based on kappa are as follows: ≤ 0 = poor, 0.01–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect.

For categorical items with only two responses available, RTI only calculated unweighted kappas. For items with more than two responses, RTI calculated both weighted and unweighted kappas. Unweighted kappa assumes the same “distance” between every one-unit difference in response across an ordinal scale (e.g., for the CARE functional item scale range 1–6, an unweighted kappa assumes the difference in functional ability between a score of 1 - dependent and 2 - substantial/maximal assist is the same as the difference in functional ability between 5 - setup or clean-up assistance and 6 - independent). RTI used Fleiss-Cohen weights, or quadratic weights, which approximate the intra-class correlation coefficient and are commonly used for calculating weighted kappas. This choice of weighting is consistent with prior analyses of assessment reliability where the method for developing weights was specified (Hirdes et al., 2002; Streiner & Norman, 1995). Note that Fleiss-Cohen weights put lower emphasis on disagreements between responses that fall near each other on an item scale. It should also be noted that the value of kappa can be influenced by the prevalence of the outcome or characteristic being measured. If the outcome or characteristic is rare, the kappa will be low because kappa attributes the majority of agreement among raters to chance. Kappa is also influenced by bias, and if the effective sample size is small, variation may also play a role in the results. Hence, we report both weighted and unweighted kappas to give the range of agreement found under the two sets of assumptions.

Additionally, RTI calculated a separate set of kappa statistics (unweighted and weighted where applicable) for items where additional responses outside of an ordinal scale were available (letter codes) and were set to missing. For example, for function items, clinicians could choose between five and six different letter codes designating that an item was “not attempted.” RTI reported a set of kappas for these items where the “not attempted” responses were recoded to missing.

**Item Reliability - Videotaped Standardized Patients Interrater Reliability:** Two main analytic approaches were used for assessing the video reliability of the CARE items, adhering closely to the methods used by Fricke, Unsworth, and Worrell (1993) in their video reliability study of the FIM® Instrument. First, percent agreement with the mode response was calculated for each of the CARE function items included in at least one of the nine videos. Unlike the approach used by Fricke et al., RTI did not consider agreement at one response level above and below the mode, and instead used a stricter approach looking at direct modal agreement only. In the second approach, percent agreement with the internal clinical team’s consensus response was also calculated. This second measure not only gives an indication of item reliability, but also reflects training consistency for the providers.

**Scale Level - Internal Consistency:** In addition to item-level reliability testing, we examined internal consistency, which provides a general assessment of how well the items interrelate within a domain or subscale. Internal consistency is assessed using the Cronbach’s alpha coefficient, which is the average correlation of all possible half-scale divisions. Cronbach’s alpha is a statistic frequently assessed when instrument or scale psychometrics are published. The Cronbach’s alpha reliability estimate ranges from zero to one, with an estimate of zero indicating that there is no consistency of measurement among the items, and one indicating perfect consistency. Many cutoff criteria exist to determine whether or not a scale shows good consistency or whether the items “hang together” well. The consensus is that Cronbach’s alpha should be at least 0.70 for an adequate scale for group-level decisions, and alphas closer to 1 indicate a good scale (Aron and Aron, 1999).

**Citations:**

Aron, A., & Aron, E. N. (1999). *Statistics for psychology* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Fricke, J., Unsworth, C., & Worrell, D. (1993). Reliability of the Functional Independence Measure with occupational therapists. *Australian Occupational Therapy Journal,* 40*,* 7–15.

Hirdes, J. P., Smith, T. F., Rabinowitz, T., Yamauchi, K., Pérez, E., Telegdi, N. C., … Fries, B. E. (2002). The Resident Assessment Instrument–Mental Health (RAI-MH): Inter-rater reliability and convergent validity. *Journal of Behavioral Health Services and Research, 29,* 419-432.

Streiner, DL, Norman, GR. (1995). *Health measurement scales: A practical guide to their development and use.* Oxford, UK: Oxford University Press.

**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*) *Data Element Reliability*

**Item Reliability - Traditional Interrater Reliability:** For the traditional reliability study, kappa statistics indicated substantial agreement among raters. The weighted kappa values for the self-care items range between 0.798 for eating to 0.869 for upper-body dressing. Unweighted kappas ranged from 0.598 for oral hygiene to 0.634 for upper-body dressing. Provider-specific analyses of self-care items show similar agreement to the overall estimates. The lower-body dressing item had the highest overall weighted kappa (0.855), while the eating item had the lowest (0.798). Unweighted overall kappas ranged from 0.636 (toileting) to 0.598 (oral hygiene). The report “The Development and Testing of the Continuity Assessment Record and Evaluation (CARE) Item Set: Final Report on Reliability Testing Volume 2 of 3” provides detailed results for the traditional interrater reliability analyses.

**Item Reliability - Videotaped Standardized Patients Reliability Study:** The video reliability study indicated substantial agreement with the mode and clinical team among all items, typically upwards of 70%. The notable exception to this trend exists among the clinicians in the “Other” category (mostly licensed professional nurses [LPNs]); they consistently had the lowest levels of agreement among all core self-care items, ranging from 50 to 72%. For the toileting and dressing items, the agreement with the clinical team was lower than with the mode. This occurred because the clinical team response differed from the mode for these three items in either one or two videos. Nonetheless, because the clinical team response and mode were identical on most of the videos, agreement was still quite high for these items. In general, study clinicians had responses on average that agreed with the expert clinical team or were slightly lower. The report “Continuity Assessment Record and Evaluation (CARE) Item Set: Video Reliability Testing” provides detailed results for the video reliability analyses.

**Scale Reliability: Internal Consistency:** Assessments of the self-care subscale at both admission and discharge tend to show good reliability statistics (Cronbach’s alpha of at least 0.80). **Table 1** provides the findings from the Cronbach’s alpha internal consistency estimates by provider type and shows that the functional status items maintain a very high internal consistency. In addition, no one provider type appears to have reliability estimates higher or lower than the rest, indicating similarity of CARE usage with respect to internal consistency.

**Table 1. CARE Functional Status Internal Consistency Reliability Summary by Provider Type**

| **CARE analytic set** | **Overall alpha** | **Home Health Agency  alpha** | **Skilled Nursing Facility  alpha** | **Inpatient Rehabilitation Facility  alpha** | **Long-Term Care Hospital  alpha** |
| --- | --- | --- | --- | --- | --- |
| Self-Care | 0.96 | 0.94 | 0.95 | 0.95 | 0.96 |

CARE = Continuity Assessment Record and Evaluation.

**Citations:**

Gage, B., Smith, L., Ross, J., et al. (2012). The Development and Testing of the Continuity Assessment Record and Evaluation (CARE) Item Set: Final Report on Reliability Testing Volume 2 of 3. RTI International. (Available at: <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/CARE-Item-Set-and-B-CARE.html>)

Smith, L., Deutsch, A., Barch, D., et al. (2012). Continuity Assessment Record and Evaluation (CARE) Item Set: Video Reliability Testing. RTI International. (Available at: <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/CARE-Item-Set-and-B-CARE.html>)

**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?*)

We used several methods to examine reliability of the items and the scale (the items as a group). Overall, we found good reliability at the item and scale levels.

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
X **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

X **Empirical validity testing**  **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

**Content Validity: Similarity of items across other self-care assessment:** We compared the items included on other functional assessment instruments with the CARE function items included in this self-care quality measure.

**Scale Reliability and Construct Validity - Item Difficulty Ordering (Rasch Analysis):** A statistical procedure called Rasch analysis was used to examine the self-care items. Rasch analysis estimates the placement of item scores along a “ruler” ordered from easy to difficult. The ordering of items from easy (bottom) to difficult (top) provides the analysis-established item difficulty hierarchy. This hierarchy can be evaluated against item design specifications (i.e., the intended construction of the items to be easy or difficult) or against an expert clinical opinion as an indication of construct validity. If items fall into unexpected locations on the hierarchy, then the content of the items should be evaluated further and potentially modified.

**Item and Scale Reliability and Validity - Fit Assessment and Response Option Assessment (Rasch Analysis):** As part of the analysis, Rasch methodology places persons and item scores on a “ruler” it creates in order to understand how easy or challenging an item is relative to other items. In addition, Rasch analysis assesses how well the items measure the patients in the sample. Rasch analysis also uses item-level response data (i.e., scores) to determine how well a set of items work together to measure a concept. Evaluations of these item-level response categories can indicate whether or not the rating scale is working as expected. That is, patients with higher overall self-care abilities show increasing levels of independence on the self-care item scores. The Rasch model can assess items, even when some data is missing for some patients. Likewise, patients can be assessed even on a reduced set of function items, after a common item set has been established. Therefore, items can be appropriately targeted to a given patient population while maintaining the ability to compare functional ability across populations. That is, a “short form” can be used in clinical settings, but also linked to the larger set of items.

Rasch analysis of the CARE self-care items was conducted to establish the best set of items to measure self-care functioning. In the measurement of functional ability, the items used in an observational assessment must maintain a degree of integrity to ensure that the concept of interest, in this case self-care, is being appropriately measured. To that end, item-level analysis can provide information on item set suitability (or how well they work together to measure a patient’s self-care ability). Items that are potentially unsuitable (or show misfit) are noted in the Rasch analysis by fit statistics that show when unexpected responses are being produced by individual items within a set of items.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

**Content Validity: Similarity of items across other self-care assessment instruments. Table 2** lists the self-care items commonly found on functional assessment instruments. All self-care activity areas included on other assessments are included on the CARE tool.

**Table. 2. Comparison of Selected Self-Care Items from the CARE Tool and Other Functional Assessment Instruments.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Item** | **Continuity Assessment Record and Evaluation (CARE)** | **Barthel Index** | **FIM® instrument** | **Katz ADL Scale** | **Minimum Data Set (MDS)** | **Outcome and Assessment Information Set-C (OASIS-C)** |
| Eating | x | x | x | x | x | x |
| Grooming/Personal or oral hygiene | x | x | x |  | x | x |
| Toileting hygiene | x | x | x | x | x | x |
| Bathing/showering | x | x | x | x | x | x |
| Dressing |  | x |  | x | x |  |
| Dressing lower Body | x |  | x |  |  | x |
| Dressing upper body | x |  | x |  |  | x |
| Dressing: Footwear | x |  |  |  |  |  |

**Construct Validity: Item Difficulty Ordering (Rasch Analysis):** The overall expected score placement of each self-care item on the concept “ruler” made sense clinically, and is consistent with other functional assessment scales. The order of the items by difficulty level, with the hardest activity listed first, is as follows:

Footwear

Lower body dressing

Toileting hygiene

Shower/bathe self

Upper body dressing

Oral hygiene

Eating

**Figure 1** reports the item hierarchy, the patient distribution and the rating scale scores in one graphic. It shows the overall expected score placement on the self-care “ruler” for each item. This ruler, which has a range of 1 to 100, is shown at the very top and the bottom of the graphic. The item hierarchy or difficulty order, from easy (bottom) to difficult (top), is shown on the right side of the graphic. For each item presented on the right, the overall expected placement of the score options (from “1” for “dependent” to “6” for “independent”) are shown along the ruler. Each item is presented on a row and the scores begin with the most dependent (represented by the “1”) on the far-left graphic boundary and the most independent (represented by “6”) on far-right graphic boundary. Finally, the threshold between two score options is represented by a colon (:), and is where a patient has an equal chance of responding in either the higher or lower category. Use of the “ruler” allows visualization of the scores for each self-care item in relation to the scores of other self-care items. The letters at the bottom of **Figure 1** describe the distribution of people along the ruler, where “M” is the average of the sample and “S” and “T” are one and two times the standard deviation around that average, respectively. The percentile values represent the distribution of patients along the “ruler.”

**Figure 1. IRF Self-Care Items: Item Rating Scale Responses and Item Hierarchy**

0 10 20 30 40 50 60 70 80 90 100

|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----| ITEM

1 1 : 2 : 3 : 4 : 5 : 6 6 Put on/Take off Footwear

1 1 : 2 : 3 : 4 : 5 : 6 6 Lower Body Dressing

1 1 : 2 : 3 : 4 : 5 : 6 6 Toileting Hygiene

1 1 : 2 : 3 : 4 : 5 : 6 6 Shower/Bathe Self

| |

1 1 : 2 : 3 : 4 : 5 : 6 6 Upper Body Dressing

| |

1 1 : 2 : 3 : 4 : 5 : 6 6 Oral Hygiene

| |

11 : 2 : 3 : 4 : 5 : 6 6 Eating

|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----| ITEM

0 10 20 30 40 50 60 70 80 90 100

T S M S T

0 10 20 30 50 70 80 90 99 PERCENTILE

**Scale Reliability and Validity - Fit Assessment (Rasch Analysis):** The Rasch model expects consistency in response patterns. Rasch analysis produces fit statistics that show when unexpected responses are being produced by individual items within a set of items. Most of the items selected for the IRF self-care item set show the response pattern expected. The Rasch model expects the challenging items to be difficult to complete for all patients. In a similar way, patients with more self-care skills are expected to have higher scores on all items. Items that don’t seem to function this way could show misfit, or unexpected responses. When we see many unexpected responses produced by an item, we need to think more about how the item is constructed and why it might be producing such unexpected responses. There are two categories of fit, one designed more for outliers (outfit) and one designed for response unexpectedness of persons near the item’s difficulty (infit). Infit and outfit mean square values above 1.4 are considered to be unacceptably unexpected or “noisy” (Wright and Linacre, 1994) when looking at multiple-point response scales. Misfit seen near the item difficulty, or large values of infit, are concerning because they indicate noise (unexpected responses) where the item should be the most productive for measurement. Items that do not elicit the expected responses are referred to as “misfitting.”

Findings and Interpretation: Overall, the self-care item set is eliciting responses as expected. **Table 3** reports fit statistics for the self-care items and shows that there is only one item, Eating, with fit statistics outside the acceptable range. The item Eating has been found to misfit in other function scales and may be related to patients who have a swallowing impairment.

**Table 3. Fit Statistics for the Self-Care Items (n = 5,129)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Admission and Discharge** | | **Admission** | | **Discharge** | |
| Item | Infit mean square | Outfit mean square | Infit mean square | Outfit mean square | Infit  mean square | Outfit mean square |
| Eating | 1.28 | 1.64 | 1.33 | 1.42 | 1.26 | 1.61 |
| Oral Hygiene | 1.02 | 1.03 | 1.01 | 1.04 | 1.07 | 1.01 |
| Toileting Hygiene | 1.00 | 0.99 | 0.99 | 1.00 | 1.08 | 1.04 |
| Upper Dressing | 0.87 | 0.85 | 0.90 | 0.91 | 0.83 | 0.79 |
| Lower Dressing | 0.59 | 0.62 | 0.61 | 0.66 | 0.60 | 0.62 |
| Shower/Bathe Self | 0.92 | 1.17 | 0.80 | 0.88 | 1.02 | 1.20 |
| Footwear | 1.09 | 1.07 | 1.18 | 1.18 | 1.07 | 1.03 |

**Item and Scale Reliability and Validity - Response Option Assessment (Rasch Analysis):** Rasch analyses provide information on how many people were observed in a particular response category (i.e., independent to dependent) and the average ability (or skill level) of those individuals. Evaluations of observed response categories indicate that rating scale use is as expected, with patients with higher skill showing increasing levels of independence. An important step in the self-care item evaluation is to establish that the 6-point observation response scale is operating as intended. It is expected that lower ability persons would generally be observed in the more dependent categories (substantial assistance, etc.). Therefore, the average ability (or skill level) estimate associated with the more dependent response categories would be lower than those associated with the more independent response categories.

Findings and Interpretation: The evaluation of the average ability (**Table 4**) as observations progress from dependent to independent indicate that, for the self-care items, the observational scale is working as intended.

**Table 4. Distribution of Admission and Discharge Scores and Average Ability Estimate by Response Code**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | Response  Code  (score) | Number | Percent | Average Ability  (0 to 100 Scale) |
| Eating |  |  |  |  |
|  | 1 | 141 | 2 | 9.85 |
|  | 2 | 114 | 2 | 21.89 |
|  | 3 | 211 | 4 | 33.09 |
|  | 4 | 711 | 12 | 42.05 |
|  | 5 | 1,871 | 33 | 49.33 |
|  | 6 | 2,645 | 46 | 70.55 |
| Oral hygiene | |  |  |  |
|  | 1 | 197 | 3 | 8.63 |
|  | 2 | 180 | 3 | 25.47 |
|  | 3 | 432 | 8 | 35.70 |
|  | 4 | 1,084 | 19 | 45.98 |
|  | 5 | 2,128 | 37 | 52.52 |
|  | 6 | 1,718 | 30 | 79.20 |
| Toileting hygiene | |  |  |  |
|  | 1 | 954 | 17 | 29.24 |
|  | 2 | 727 | 13 | 40.58 |
|  | 3 | 945 | 16 | 48.04 |
|  | 4 | 1,378 | 24 | 55.98 |
|  | 5 | 638 | 11 | 65.92 |
|  | 6 | 1,132 | 20 | 87.18 |
| Upper body dressing | |  |  |  |
|  | 1 | 303 | 5 | 14.94 |
|  | 2 | 482 | 8 | 32.43 |
|  | 3 | 993 | 17 | 42.71 |
|  | 4 | 1248 | 22 | 51.13 |
|  | 5 | 1558 | 27 | 57.84 |
|  | 6 | 1206 | 21 | 87.10 |
| Lower body dressing | |  |  |  |
|  | 1 | 923 | 16 | 28.31 |
|  | 2 | 1133 | 20 | 42.47 |
|  | 3 | 1159 | 20 | 50.93 |
|  | 4 | 1098 | 19 | 58.94 |
|  | 5 | 575 | 10 | 69.31 |
|  | 6 | 901 | 16 | 92.92 |

(cont’d)

**Table 4. Distribution of Admission and Discharge Scores and Average Ability Estimate by Response Category (cont’d)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | Response  Code | Number | Percent | Average Ability  (0 to 100 Scale) |
| Shower/bathe self | |  |  |  |
|  | 1 | 385 | 9 | 23.37 |
|  | 2 | 621 | 14 | 38.54 |
|  | 3 | 1031 | 23 | 48.62 |
|  | 4 | 1215 | 27 | 59.86 |
|  | 5 | 663 | 15 | 71.56 |
|  | 6 | 597 | 13 | 94.95 |
| Putting on and taking off footwear | | |  |  |
|  | 1 | 1269 | 24 | 33.83 |
|  | 2 | 832 | 16 | 44.58 |
|  | 3 | 830 | 16 | 52.06 |
|  | 4 | 905 | 17 | 58.81 |
|  | 5 | 600 | 11 | 67.15 |
|  | 6 | 903 | 17 | 91.75 |

Note: Activity not attempted/did not occur codes are not included in this analysis.

\*\*Response categories are defined as: 1 – Dependent; 2 – Substantial/maximal assistance; 3 - Partial/moderate assistance; 4 - Supervision or touching assistance; 5 - Setup or clean-up assistance; and 6 - Independent.

**Citation:**

Wright BD, Linacre JM (1994) Reasonable mean-square fit values. *Rasch Measurement Transactions.* 8:3 p.370. <http://www.rasch.org/rmt/rmt83b.htm>

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The activities included in the self-scale scale are similar to self-care items on other functional assessment instruments. According to technical expert consultation, the self-care items are in an order that makes sense clinically. That is, the order of the items from easy to difficult (item hierarchy) is consistent with task difficulties in the field. The item hierarchy is also consistent with clinical findings from similar instruments. Rasch analysis of the data showed the items work well together, with generally good infit and outfit statistics. As expected, for each item, the average ability score of patients increases as the rating scale increases.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA** ☐ **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We examined the percentage of patients who were excluded from the quality measure calculations due the exclusion criteria. With the exception of the missing data criteria, the exclusion criteria are applied to the data in order to maintain the validity of the performance score.

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**Table 5** lists the exclusion criteria and the number and percent of patient records excluded for each criterion. Overall, 13.5 percent of the sample met one or more of the exclusion criteria. With the exception of the missing data criteria, the exclusion criteria are applied to the data in order to maintain the validity of the performance score. Missing data was present in less than 1 percent of patient records after all other exclusion criteria were applied.

**Table 5. Number and Percentage of Inpatient Rehabilitation Facility Patients Excluded from the Self-Care Quality Measure (n = 5,516)**

|  |  |  |
| --- | --- | --- |
| **Exclusion Criteria** | **N** | **Percent** |
|
| Incomplete stay: discharged due to medical emergency | 511 | 9.26 |
| Incomplete stay: discharged against medical advice | + | + |
| Incomplete stay: length of stay < 3 days | 73 | 1.32 |
| Discharged to hospice (institutional facility) | + | + |
| Discharged to another IRF | 16 | 0.29 |
| Patients with medical conditions with limited or less predictable improvement in self-care function with current set of items: complete quadriplegia and locked-in syndrome | + | + |
| Patients with medical conditions with less predictable improvement in self-care function with current set of items: coma, persistent vegetative state, severe brain damage at admission | 180 | 3.26 |
| No self-care limitations at admission assessment | + | + |
| Patient records with all function data missing at admission | 18 | 0.33 |
| Patient records with all function data missing at discharge | 196 | 3.55 |
| Patient records with all function data as letter codes at admission and discharge | 0 | 0 |
| Patients records with all function data missing or letter codes at admission and discharge | + | + |
| **Total number of case records excluded\*** | **747** | **13.54** |

\*More than one criterion may apply, so the total number is less than the sum of all criteria.

+ Cells based on a sample size of n < 11 are not shown.

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis. Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Overall, 13.5 percent of the sample met one or more of the exclusion criteria. With the exception of the missing data criteria, the exclusion criteria are applied to the data in order to maintain the validity of the performance score. Missing data was present in less than 1 percent of patient records after all other exclusion criteria were applied.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

☐ **No risk adjustment or stratification**

X **Statistical risk model with**  **74** **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

**2b4.2. If an outcome or resource use 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**.

Not applicable. This quality measure is risk adjusted.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient 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 and not related to disparities*)

This quality measure estimates the risk adjusted mean change in self-care score between admission and discharge among IRF patients. Functional improvement can vary based on patients’ demographic or clinical characteristics, therefore, this measure is risk adjusted. The goal of risk adjustment is to control for differences across facilities in patient characteristics at admission that might be related to the outcome of interest. This allows outcomes to be compared across facilities after differences in patient complexity (i.e., patient characteristics) have been accounted for in the analysis. The risk adjustment model for this measure controls for variation across facilities in patient demographics (e.g., age) and clinical (e.g., diagnosis) characteristics present at the time of admission that may influence functional outcomes, to allow change in self-care outcomes to be compared across IRFs.

To develop the risk adjustment model for this measure, we used three data sources that were linked: 1) CARE data from the Post-Acute Care Payment Reform Demonstration, 2) Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-PAI) data to determine the primary condition requiring IRF admission and the comorbidities, and 3) Claims data from the acute care and long-term care hospital stay prior to the IRF stay to determine comorbidities. Development and testing of the risk adjustment model was conducted after applying the exclusion criteria described in 2b3, and was based on 4,769 patient records from 38 IRFs.

***Risk Adjustor Selection – Conceptual Rationale and Statistical Testing***

An initial, extensive set of risk adjustors was selected based on literature review, clinical relevance, and findings from the Post-Acute Care Payment Reform Demonstration analyses. This initial set of risk adjustors was refined based on input from a Technical Expert Panel convened by our measure development contractor. We also requested input on risk adjustors as part of the public comment process, and tested public suggestions as part of the risk adjustment model development.

We tested the risk adjustors and developed the risk adjustment model using multiple linear regression analyses. The dependent variable was the change in self-care score for each patient, calculated as the difference between the discharge self-care score and admission self-care score. We made decisions to retain or drop each risk adjustor based on its sample size, regression coefficient, significance level, and clinical relevance to self-care outcomes. To strengthen sample sizes, when appropriate, we collapsed clinically similar risk adjustors that had low prevalence. In general, a *p*-value of 0.10 was used to determine statistical significance. However, we retained variables that approached statistical significance or those that did not reach significance if they were clinically related to self-care outcomes, or had large regression coefficients. Final risk adjustor selection was based on a combination of clinical reasoning and statistical findings.

Once we determined the final set of risk adjustors using ordinary least squares multiple linear regression, we ran a generalized linear model using generalized estimation equations (GEE) as the estimation method to account for clustering of data within each IRF. The generalized estimation equations method accounted for potentially correlated outcomes of patients within the same IRF, in addition to risk adjusting the change in self-care outcome using the final set of risk adjustors.

Risk adjustors included in the final model are described below, and also presented in Attachment 1.

***Age Groups:*** We included seven age groups in the risk adjustment model (< 35 years, 35–44 years, 45–54 years, 55–64 years, 75–84 years, 85–90 years, and ≥ 90 years). The age group 65–74 years formed the reference category. Age was not normally distributed in our sample, so it was more appropriate to use age groups in our analyses. When compared to the reference group (patients 65–74 years), patients younger than 35 years had significantly smaller change in self-care scores (coefficient = -2.216, *p* = 0.0758). Patients 75–84 years (coefficient = -0.651, *p* = 0.0013), 85–90 years (coefficient = -1.528, *p* < 0.0001), and over 90 years (coefficient = -2.246, *p* < 0.0001) also had significantly, and progressively, smaller change in self-care scores than patients in the reference category. Patients 35–44 years, 45–54 years, and 55–64 years did not have significantly different change scores compared with the reference category. Nevertheless, we chose not to collapse these groups based on public comment feedback regarding the clinical importance of maintaining fine discrimination among age groups.

***Admission Self-Care Scores:*** Since improvement in self-care during the IRF stay may vary based on admission self-care ability, we risk adjusted for admission self-care scores in our regression model. Scatter plots of admission self-care scores and change in self-care scores showed a non-linear relationship between the two variables. Therefore, we included admission self-care scores in two forms in the model: a continuous form, and a squared form to account for the curvilinear relationship. The squared form of admission self-care scores was significant (coefficient = -0.016, *p* < 0.0001), while the continuous form did not have a significant effect (coefficient = 0.044, *p* = 0.6663). Examination of overall factor significance results, however, showed that admission self-care was a significant predictor (*p* < 0.0001), indicating that the non-significant admission self-care effect noted was related to the choice of reference categories and model intercept. Thus, we included both forms of admission self-care in the final model.

***Primary Diagnosis Groups Based on IRF Primary Diagnosis:*** The expert panel members agreed that medical diagnosis was an important risk adjustor, and they unanimously recommended using the IRF primary diagnosis rather than the primary diagnosis from the prior acute care stay in the risk adjustment model. We used Impairment Group codes reported on the IRF-PAI (Item 21) to create the following 13 mutually-exclusive primary diagnosis groups: (1) stroke, (2) non-traumatic brain dysfunction, (3) traumatic brain dysfunction, (4) non-traumatic spinal cord dysfunction, (5) traumatic spinal cord dysfunction, (6) progressive neurological conditions, (7) other neurological conditions (for e.g., polyneuropathy), (8) fractures and other multiple trauma, (9) hip and knee replacements, (10) amputation, (11) other orthopedic conditions (for e.g., arthritis), (12) debility and cardiorespiratory conditions, and (13) medically complex conditions. “Hip and knee replacements” formed the reference category, and the remaining 12 primary diagnosis groups were risk adjustors in the model. When compared to the reference category, all diagnosis groups were significant predictors of change in self-care scores, with the exception of the “non-traumatic brain dysfunction” and “medically complex conditions” groups. The remaining ten primary diagnosis groups had significantly smaller change in self-care scores compared with the “hip and knee replacements” group. The “traumatic spinal cord dysfunction” group had the largest coefficient (-10.715, *p* < 0.0001).

***Interaction between Primary Diagnosis Groups and Admission Self-Care Scores:*** To account for the possibility that the relationship between admission self-care and change in self-care scores may vary based on the patient’s diagnosis group, we tested interaction terms between admission self-care scores (continuous form) and each diagnosis group included in the model. Thus, 12 interaction terms for admission self-care by diagnosis group were tested. Nine interaction terms were significant, as shown in Attachment 1. All interaction terms were retained in the model.

***Prior Acute or Long-Term Care Hospital Primary Diagnosis (Surgical or Medical):*** We classified primary diagnosis from the prior acute or long-term care hospital stay as surgical or medical. We included this indicator of surgical or medical diagnosis in the model, with the “surgical” category being the risk adjustor and the “medical” category, the reference group. Patients with a surgical diagnosis in the prior acute or long-term care hospital had significantly larger change in self-care scores (coefficient = 0.755, *p* = 0.0002) compared with the reference group.

***Prior Functioning - Self-Care:*** We included patients’ functional ability in self-care before onset of their current illness, injury or exacerbation, as a risk adjustor in the model. We included separate categories for patients who were “dependent”, and those who needed “some help” in self-care before their current illness. Patients who were “previously independent” in self-care formed the reference category. Patients who were previously “dependent” in self-care and those who needed “some help” had significantly smaller change scores compared with the reference category. The negative coefficient for the “dependent” category (-1.894, *p* = 0.0032) was larger than that for the “some help” (-0.668, *p* = 0.0117) category.

***Prior Functioning - Indoor Ambulation:*** We included patients’ functional ability in indoor ambulation before onset of their current illness, injury or exacerbation, as a risk adjustor in the model. We combined the dependent and some help categories into one group, and patients who were previously independent in indoor ambulation formed the reference category. Patients who were previously dependent and those who needed some help in indoor ambulation had significantly smaller change in self-care scores (coefficient = -1.026, *p* = 0.0008) compared with the reference category.

***Prior Mobility Devices/Aids:*** We risk adjusted for use of four types of mobility devices or aids before the current illness, injury, or exacerbation, including walker, wheelchair/scooter, mechanical lift, and orthotics or prosthetics. Prior use of each of these mobility devices or aids was associated with significantly smaller change in self-care scores, with prior use of a mechanical lift having the largest coefficient (-4.307, *p* = 0.0001), followed by prior use of orthotics or prosthetics (coefficient = -1.673, *p* = 0.0809).

***Stage 2 Pressure Ulcer****:* Our risk adjustment model included an indicator variable for the presence of one or more stage 2 pressure ulcers on admission, with the reference category being patients who did not have a stage 2 pressure ulcer. Patients with stage 2 pressure ulcers had a significantly smaller change in self-care scores (coefficient = -1.200, *p* = 0.0003) compared with the reference category.

***Stage 3, 4, or Unstageable Pressure Ulcers:*** We included an indicator variable for the presence of one or more stage 3, 4, or unstageable pressure ulcers, with the reference category being patients who did not have such ulcers. Patients with stage 3, 4, or unstageable pressure ulcers had significantly smaller change in self-care scores (coefficient = -1.230, *p* = 0.0561) compared with the reference category.

***Cognitive Function Assessed by the Brief Interview for Mental Status:*** Based on Brief Interview for Mental Status scores, patients’ cognitive function was classified as intact or borderline, moderately impaired, or severely impaired. “Moderately impaired” and “severely impaired” cognitive function were included as two separate risk adjustors in the model, while “intact or borderline” cognitive function formed the reference category. Patients with moderately impaired cognitive function (coefficient = -0.809, *p* = 0.0002) and those with severely impaired cognitive function (coefficient = -2.179, *p* < 0.0001) had significantly smaller change scores compared with the reference category.

***Communication Impairment:*** Communication impairment includes both expression (expression of ideas and wants) and comprehension (understanding verbal content) abilities. While expression and comprehension abilities are separate assessment items, we combined them into a single communication impairment risk adjustor given these two variables were correlated, with considerable overlap in patients who had expression and comprehension impairment and based on input from the expert panel. The final risk adjustment model included “moderate to severe communication impairment*”* as a risk adjustor, with this group having significantly smaller change in self-care scores (coefficient = -1.449, *p* < 0.0001) compared with the reference category of “mild or no communication impairment.”

***Bladder Incontinence:*** We included a single risk adjustor for bladder incontinence, which comprises patients with bladder incontinence “less than daily,” “daily,” or “always.” The reference category included patients who had “stress incontinence only, were always continent, or had no urine output”. Patients with bladder incontinence (coefficient = -0.923, *p* = 0.0002) had significantly smaller change in self-care scores compared with the reference category. We also included a risk adjustor for patients with an indwelling urinary catheter (coefficient = -0.731, *p* = 0.0027).

***Bowel Incontinence:*** We included two separate risk adjustors related to bowel incontinence: “always incontinent” and “less than daily or daily incontinence.” The reference category included patients who “were always continent, had no bowel output during the assessment period, or had a bowel catheter management system”. Patients with bowel incontinence had significantly smaller change in self-care scores compared with the reference group, with the “always incontinent” category (coefficient = -1.582, *p* = 0.0021) having a larger negative coefficient compared with the “less than daily” or “daily incontinence” category (coefficient = -0.635, *p* = 0.0241).

##### *Swallowing Ability:* Our model included two separate risk adjustors related to patients’ swallowing ability: (1) need for modified food consistency or supervision, and (2) need for tube or parenteral feeding. Both risk adjustors were significantly predictive of smaller change in self-care scores, with the tube or parenteral feeding group (coefficient = -1.056, *p* = 0.0864) having a larger coefficient than the modified food consistency or supervision group (coefficient = -0.562, *p* = 0.0584).

***Comorbidities:*** We used all the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes reported on the IRF-PAI (Item 24 - Comorbid Conditions), and on prior acute claims within the past 100 days to identify patient comorbidities. ICD-9-CM codes were used to assign patients into one or more of the 201 Hierarchical Condition Categories. We tested approximately 120 of the Hierarchical Condition Categories that were determined to be clinically relevant to self-care outcomes. We collapsed several similar Hierarchical Condition Categories into groups during model development due to the low prevalence of some conditions in the sample. Grouping of Hierarchical Condition Categories was based on clinical similarity as well as similarity of regression estimates.

To ensure that the same diagnoses or conditions were not represented in both the primary diagnosis groups and comorbidities, we applied exclusion criteria such that certain comorbidities were excluded if they were also present as primary diagnoses. For example, tetraplegia and paraplegia were excluded as comorbidities if the patient’s primary diagnosis group was “non-traumatic spinal cord dysfunction” or “traumatic spinal cord dysfunction”; amputation was excluded as a comorbidity if the patient’s primary diagnosis group was “amputation.”

Attachment 1 shows the regression coefficients and significance values for all comorbidities in the final risk adjustment model. We retained comorbidities that were clinically important or had large coefficients, even when they were not statistically significant. Comorbidities with the largest negative coefficients, indicating smaller change in self-care scores, include central nervous system infections; certain cancers; dementia; tetraplegia; multiple sclerosis; angina pectoris; hypertensive heart disease; kidney transplant status; and dialysis and stage 5 chronic kidney disease.

***Model for Individual Patient’s Expected Change in Self-Care Score***

We used results from the generalized linear model to calculate the expected change in self-care score for each patient using the formula below:

*Expected change in self-care score =*

intercept + (age group \* coefficient) + (admission self-care \* coefficient) + (squared admission self-care \* coefficient) + (primary diagnosis group \* coefficient) + (interaction term coefficient for admission self-care and primary diagnosis group) + (surgical or medical prior acute or LTCH primary diagnosis \* coefficient) + (prior functioning: self-care \* coefficient) + (prior functioning: indoor ambulation \* coefficient) + (prior use of walker \* coefficient) + (prior use of wheelchair \* coefficient) + (prior use of mechanical lift \* coefficient) + (prior use of orthotics/prosthetics \* coefficient) + (stage 2 pressure ulcer \* coefficient) + (stage 3, 4 or unstageable pressure ulcer \* coefficient) + (cognitive function\* coefficient) + (communication impairment \* coefficient) + (bladder incontinence \* coefficient) + (bowel incontinence \* coefficient) + (swallowing ability: modified food consistency or supervision \* coefficient + (swallowing ability: tube/parenteral feeding \* coefficient) + (comorbidity \* coefficient)

In the equation above, the intercept and coefficient values were constant for each patient, while risk adjustor values were specific to the patient. For patients with multiple comorbidities, each comorbidity was included in the equation.

***Risk Adjusted Change in Self-Care Outcome for Each IRF***

To calculate the risk adjusted change score for each IRF, we first computed three values:

##### Mean observed change in self-care score for each IRF: We calculated the mean observed change score for each IRF as the mean of the observed change in self-care scores for all patients in the IRF.

##### Mean expected change in self-care score for each IRF: As described above, we modeled each patient’s expected change in self-care score using results from the generalized linear model. We then computed the mean expected change in self-care score for each IRF by calculating the mean of the expected change score for all patients in the IRF.

##### National mean observed change in self-care score: We calculated national mean observed change in self-care score using data for all patients and all IRFs.

Using the above three values, the risk adjusted change in self-care outcome for each IRF was calculated using the formula:

ICD-10 coding: For this quality measure, we use ICD-9-CM codes to identify patients with comorbidities. More specifically, we use the ICD codes to assign patients into Hierarchical Condition Categories. RTI International has a draft crosswalk of ICD-10 codes that map to the Hierarchical Condition Categories, which can be used for this quality measure when ICD-10 codes are implemented.

2b4.4. What were the statistical results of the analyses used to select risk factors?

##### Results of the final risk adjustment model are shown in Attachment 1, along with regression coefficients and significance values of the final set of risk adjustors. The overall model was a significant predictor of change in self-care scores, with a *p*-value less than 0.0001. The overall model R-square was 0.2578, indicating that 25.78% of the variance in change in self-care was explained by the model. In general, regression coefficients of individual risk adjustors demonstrated that the predictive ability of risk adjustors was as clinically expected. Examination of Type II semi-partial correlation coefficients indicated that cognitive function and primary diagnosis group were the strongest predictors, explaining 1.20% and 1.10% of the variance in change in self-care, respectively.

Distributions of the facility-level mean unadjusted and risk adjusted change in self-care scores are shown in **Figures 2 and 3 and Table 6.** Important differences in the distribution of the two sets of scores were noted, which speaks to the importance of risk adjustment. **Figure 2** shows that the facility-level mean unadjusted change scores are largely concentrated in the center of the distribution, with fewer IRFs at the extremes of the distribution, particularly at the higher extreme. In contrast, **Figure 3** demonstrates normal distribution and good variability of the facility-level mean risk adjusted change in self-care scores.

**Figure 2. Distribution of Facility-Level Mean Unadjusted Change in Self-Care Scores**

**Figure 3. Distribution of Facility-Level Mean Risk Adjusted Change in Self-Care Scores**

Skewness and kurtosis values of the facility-level mean unadjusted change scores are larger than those of the mean risk adjusted change scores (**Table 6**), indicating that the unadjusted scores deviate from a normal distribution to a larger extent than the risk adjusted scores. **Table 6** shows that the standard deviation and standard error of the mean risk adjusted change scores are slightly smaller than those of the unadjusted change scores. The mean risk adjusted change scores have a range of 6.53 to 14.78, and an interquartile range of 2.07. In contrast, the mean unadjusted change scores have a wider range of 6.35 to 16.81, and a narrower interquartile range of 1.87.

**Table 6. Distribution of Facility-Level Mean Unadjusted and Risk Adjusted Change in Self-Care Scores**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Change in Self-Care Score** | **N** | **Mean (SD)** | **SE** | **Min** | **10th Pctl** | **25th Pctl** | **Median** | **75th Pctl** | **90th Pctl** | **Max** | **Skewness** | **Kurtosis** |
| Unadjusted (Observed) | 38 | 10.13 (2.07) | 0.34 | 6.35 | 8.07 | 8.98 | 10.02 | 10.85 | 11.18 | 16.8 | 1.07 | 2.93 |
| Risk Adjusted | 38 | 10.02 (1.72) | 0.28 | 6.53 | 8.06 | 8.89 | 9.82 | 10.96 | 12.34 | 14.78 | 0.50 | 0.42 |

N = Number; SD = Standard deviation; SE = standard error; Min = Minimum; Pctl = Percentile; Max = Maximum

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

Our risk adjustment model demonstrates reasonable predictive validity for IRF change in self-care scores. Using multiple linear regression, we conducted regression diagnostics to assess model performance, examining predictive ability, multicollinearity, and outlier influence. Overall, the model explained 25.78% of variance in change in self-care. The unexplained variance may be attributed to variation in facility characteristics or quality of care, or to patient characteristics not included in the model.

We assessed for multicollinearity in the final risk adjustment model based on variance inflation factor values, with values of 10 or higher suggesting multicollinearity. A high degree of multicollinearity indicates that the risk adjustors are highly associated with each other, and threatens the accuracy of individual risk adjustor coefficients. All variance inflation factor values in our model were well below 10 (range = 1.012 to 1.959), indicating that multicollinearity was not a concern. We conducted outlier influence analysis to assess for any outlying observations that may have large or extreme effects on the change in self-care outcome, with a Cook’s D score of 1.0 or higher suggesting a potentially influential observation. All Cook’s D scores were less than 1.0, with the maximum score being 0.014.

To assess model performance and stability across the sample, we divided our dataset into deciles of predicted values and calculated the ratio of average predicted change score to average observed change score within each decile. A ratio of 1 would indicate perfect agreement between average predicted and observed change scores. As seen in **Table 7**, the average predicted to observed change score ratios within each decile approximated 1.0, with a range of 0.934 and 1.055, validating model performance. Additionally, there was little variability in average predicted to observed change score ratios across deciles, supporting model stability across the range of predicted change scores and across the sample.

**Table 7. Ratio of Average Predicted to Average Observed Change in Self-Care Scores Across Deciles of Predicted Change Scores**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Deciles of Predicted Change Scores** | **Sample size** | **Average Predicted Change Score** | **Average Observed Change Score** | **Average Predicted to Observed Ratio** |
| 1 | 476 | 15.17 | 15.61 | 0.972 |
| 2 | 477 | 13.30 | 13.14 | 1.012 |
| 3 | 477 | 12.27 | 12.02 | 1.021 |
| 4 | 477 | 11.44 | 11.45 | 0.999 |
| 5 | 477 | 10.65 | 10.55 | 1.010 |
| 6 | 477 | 9.82 | 9.31 | 1.055 |
| 7 | 477 | 8.93 | 9.05 | 0.987 |
| 8 | 477 | 7.92 | 7.79 | 1.017 |
| 9 | 477 | 6.51 | 6.25 | 1.042 |
| 10 | 477 | 3.78 | 4.05 | 0.934 |
| **Total Sample** | **4,769** | **9.98** | **9.92** | **1.006** |

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*):

See Attachment 1.

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Not applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

**2b4.9. Results of Risk Stratification Analysis**:

Not applicable

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

In summary, our results demonstrate reasonable predictive ability of our risk adjustment model for IRFs.

**2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

None.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*   
 No analyses conducted.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)  
Not applicable.

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)  
Not applicable.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

**Note***: This criterion is directed 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).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of 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*)  
Not applicable

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)  
Not applicable

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of 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*)  
Not applicable

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)  
 We ran frequencies of missing data for each self-care item.

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

The frequencies of missing data for each self-care item using data from the Post-Acute Care Payment Reform Demonstration are reported in **Table 8** at admission, discharge and combined. Overall, less than one percent of data was missing.

**Table 8. Frequencies of Missing Data: IRF Patients and IRF Self-Care Quality Measure items**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **IRF Admission and Discharge**  **(n = 10,258)** | | **IRF Admission**  **(n = 5,129)** | | **IRF Discharge**  **(n = 5,129)** | |
|  | **Number Missing** | **Percent Missing** | **Number Missing** | **Percent Missing** | **Number Missing** | **Percent Missing** |
| Eating | 28 | 0.27 | + | + | 18 | 0.35 |
| Oral hygiene | 28 | 0.27 | + | + | 18 | 0.35 |
| Toileting hygiene | 28 | 0.27 | + | + | 18 | 0.35 |
| Shower/bathe self | 49 | 0.48 | 18 | 0.35 | 31 | 0.60 |
| Upper body dressing | 28 | 0.27 | + | + | 18 | 0.35 |
| Lower body dressing | 28 | 0.27 | + | + | 18 | 0.35 |
| Putting on/taking off footwear | 49 | 0.48 | 18 | 0.35 | 31 | 0.60 |

+ Cells based on a sample size of n < 11 are not shown.

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

As previously noted, we recoded ‘Activity not attempted’ codes and missing data to ‘1 – Dependent.’ This approach is consistent with the approach used with IRF functional status data published in the literature. Across all items, the percentage of data that was missing was less than one percent. This very small percentage is unlikely to cause significant bias.