**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 2633

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

**Date of Submission**: 1/7/2019

**Type of Measure:**

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

|  |
| --- |
| **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, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 25 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). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing 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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **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 **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.  **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**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)  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome 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.   **2b4.** 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.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** 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. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*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: Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-PAI) | other: Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-PAI) |

**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 primary dataset used for calculating this performance measure was the National IRF-PAI data. A copy of the IRF-PAI can be found on the following website: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-PAI-and-IRF-QRP-Manual.html>

We used two additional data sources for measure testing only to provide facility and patient-level characteristics not available in the IRF-PAI. These sources are not used for quality measure calculation:

For analyses that involved facility characteristics, we used the Provider of Service file:

* **Provider of Services Current Files (POS File):** We used the POS file to describe the characteristics of IRFs, such as census region, ownership type, and rurality, reported in **Table 1**. The POS file contains data on characteristics of hospitals and other types of healthcare facilities, including the name and address of the facility and the type of Medicare services the facility provides, among other information. The data are collected through the CMS Regional Offices. General information about the POS Files is available at: <https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Provider-of-Services/index.html>.

As described in more detail below in section 2b3.4b., this performance measure does not adjust for social risk factors. However, we have conducted testing of social risk factors, and for this testing, we used data from the Integrated Data Repository (IDR) file to capture patients’ dual eligibility status. We extracted dual eligibility data from the IDR and added this variable to our primary dataset, the IRF-PAI:

* **Beneficiary Fact table (V2\_MDCR\_BENE\_FCT) from the Integrated Data Repository (IDR):** CMS maintains the Integrated Data Repository (IDR), a high-volume data warehouse integrating Parts A, B, C, D, and DME claims, beneficiary and provider data sources, along with ancillary data such as contract information and risk scores.
* We used the IDR file to extract information on beneficiary dual eligibility status for social risk factor testing. These data are submitted by states to CMS and provide a monthly snapshot representing beneficiary characteristics as of set points in time. We used the BENE\_DUAL\_STUS\_CD (Beneficiary Point of Sale Dual Status Code) that identifies the entitlement status for the dual eligible beneficiary. Missing data is rare and if it is missing for one month’s data then the months before and after can be used. In this analysis, missing data for dual eligibility occurred for < 11 patient stays. General information about the IDR is available at: <https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/IDR/>.

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

For most testing reported in this document, we analyzed the records of patients discharged in calendar year 2017 (January 1, 2017 through December 31, 2017; 12 Months). For the Rasch analysis and internal consistency testing, we analyzed the records of patients discharged in fiscal year 2017 (October 1, 2016 through September 30, 2017; 12 Months).

**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.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: 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*)

**Inpatient Rehabilitation Facilities Included in the National IRF-PAI Data** **- Calendar Year 2017 Data**

Testing for this performance measure involved several types of data element, scale/instrument and computed performance score reliability and validity analyses, performance measure variability analyses, and social risk factor analysis. The unit of analysis for the data element and scale/instrument analyses is patient assessments or patient stays, and the unit of analysis for the computed performance measure score analyses is providers (i.e., IRFs). National data collection for the change in self-care functional status outcome measure began October 1, 2016 with the 2016 release (Version 1.4) of the IRF–Patient Assessment Instrument (IRF-PAI).

A total of 1,129 IRFs submitted IRF-PAI records during the testing period, January – December 2017. This represents 100% of this type of provider as defined by the Centers for Medicare and Medicaid Services.

**Table 1** displays the geographical location and facility characteristics of IRFs that reported IRF-PAI data for this performance measure. The majority of these IRFs are located in the southern United States (CMS Regions 4, 5, and 6) with over 20 percent in Region 6 (TX, LA, AR, OK, NM). The majority of IRFs are in urban settings (86.4%) and under private ownership (56.7%). About 25 percent of IRFs are rehabilitation hospitals; most IRFs are units. Few IRFs are teaching facilities (12.1%). Facility size is presented based on the number of patient stays. Approximately 50 percent of facilities treated 296 or fewer patients who were discharged in 2017, and the range was one stay to 4,416 patient stays. Note that providers with less than 20 stays during the 12-month testing period are excluded from facility-level analyses presented below.

Table 1. Number of IRFs Reporting by Facility Characteristics, Calendar Year 2017 (N=1,129)

| **Characteristic** | **Number (Percent)** |
| --- | --- |
| CMS Region |  |
| Region 1: CT, ME, MA, NH, RI, VT | 34 (3.0%) |
| Region 2: PR, VI, NY, NJ | 71 (6.3%) |
| Region 3: MD, DC, DE, WV, VA, PA | 122 (10.8%) |
| Region 4: NC, SC, TN, FL, GA, AL, KY, MS | 197 (17.5%) |
| Region 5: MI, MN, OH, IL, IN, WI | 209 (18.5%) |
| Region 6: TX, LA, AR, OK, NM | 233 (20.6%) |
| Region 7: MO, KS, IA, NE | 75 (6.6%) |
| Region 8: ND, UT, SD, WY, CO, MT | 43 (3.8%) |
| Region 9: NV, AZ, CA, HI, AS, Pacific Territories | 113 (10.0%) |
| Region 10: WA, AK, ID, OR | 32 (2.8%) |
| Urbanicity |  |
| Rural | 154 (13.6%) |
| Urban | 975 (86.4%) |
| Ownership Type |  |
| Government | 119 (10.5%) |
| Private | 640 (56.7%) |
| Non-profit | 370 (32.8%) |
| Rehabilitation hospital | 281 (24.9%) |
| Teaching Facility | 137 (12.1%) |
| Number of Patient Stays |  |
| Decile 1: 1-104 | 125 (11.1%) |
| Decile 2: 105-152 | 114 (10.1%) |
| Decile 3: 153-192 | 113 (10.0%) |
| Decile 4: 193-240 | 108 (9.6%) |
| Decile 5: 241-296 | 112 (9.9%) |
| Decile 6: 297-361 | 112 (9.9%) |
| Decile 7: 362-480 | 112 (9.9%) |
| Decile 8: 481-694 | 111 (9.8%) |
| Decile 9: 695-1,022 | 111 (9.8%) |
| Decile 10: 1,024-4,416 | 111 (9.8%) |

Note: Values are reported as frequency (percent); Source: RTI analysis of IRF-PAI January – December 2017, and Provider of Service (POS) File 2017 (Program reference: LP57)

**Rasch Analysis Sample using National IRF-PAI Data – Fiscal Year 2017 Data**

As noted above, the reliability and validity testing that involved Rasch analysis and internal consistency testing was conducted using fiscal year 2017 data. This dataset included 1,126 IRFs. The characteristics of these IRFs are very similar to the provider data for the calendar year 2017 data reported above.

**Face Validity – Technical Expert Panel (TEP) Survey**

On March 27, 2017, RTI International, on behalf of the Centers for Medicare & Medicaid Services (CMS), convened an in-person Technical Expert Panel (TEP) in Baltimore, MD, to seek expert input on the Inpatient Rehabilitation Facilities Quality Reporting Program (IRF QRP) quality measures, including the functional status performance measures. A pre-TEP survey completed by 7 of the 10 TEP members provided us with some data to address face validity of the Change in Self-Care performance measure. The entities that the 10 TEP members represented were: 30% non-profit organization, 40% for-profit corporations, 20% government entities, and 10% professional association. Four of the TEP members have academic affiliations. The TEP members reported their residence in the following states: Alabama, California, Kentucky, Massachusetts, Minnesota, New York, North Carolina, Ohio, Texas.

**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*)

**Total Number of Patients Included in the National IRF-PAI Data - Calendar Year 2017 Data**

IRFs submitted a total of 493,209 patient records for Medicare Part A and Medicare Advantage patient stays that ended during the testing time period (January 1 through December 31, 2017). The sociodemographic and stay-level characteristics of these Medicare patients are summarized in **Table 2**.

Patients older than the age of 65 accounted for nearly 87 percent of IRF patients. Female patients comprised just over half of the patients, nearly 80 percent of patients were white, and just under half were married. Overall, most patients lived with family or relatives prior to their IRF stay (65.4%) and more than 90 percent were admitted to the IRF from short-term general acute care hospitals. Stroke was the largest primary diagnosis group (23.3%) with debility and cardiorespiratory conditions (17.4%), fractures and other multiple trauma (11.7%), and other neurological conditions other than progressive neurological conditions (11.4%) as other major primary conditions. The majority of IRF patient stays ended with the patient discharged to home with or without care from a home health service organization (73.9%). About 15 percent of patients were discharged to other post-acute care settings, and 10 percent were discharged to a short-term general acute care hospital.

Table 2. IRF Medicare Patient and Stay Characteristics, Patients Discharged in Calendar Year 2017 (N=493,209)

| **Characteristic** | **Number (Percent)** |
| --- | --- |
| Age |  |
| 64 and younger | 66,395 (13.5%) |
| 65 to 74 | 169,773 (34.4%) |
| 75 to 84 | 161,473 (32.7%) |
| 85 and older | 95,568 (19.4%) |
| Gender |  |
| Male | 231,751 (47.0%) |
| Female | 261,458 (53.0%) |
| Race/Ethnicity\* |  |
| White | 390,837 (79.2%) |
| Black or African American | 54,971 (11.2%) |
| Hispanic or Latino | 23,361 (4.7%) |
| Asian | 7,876 (1.6%) |
| American Indian/Alaskan Native | 1,724 (0.4%) |
| Native Hawaiian/Pacific Islander | 1,954 (0.4%) |
| Marital Status |  |
| Married | 231,146 (46.9%) |
| Widowed | 131,663 (26.7%) |
| Other\*\* | 130,400 (26.4%) |
| Pre-Hospital Living With |  |
| Living Alone | 1443,592 (29.1%) |
| Family/Relatives | 322,605 (65.4%) |
| Other\*\*\* | 27,012 (5.5%) |
| Primary Diagnosis |  |
| Stroke | 114,722 (23.3%) |
| Hip or knee replacement | 20,882 (4.2%) |
| Non-traumatic brain dysfunction | 36,147 (7.3%) |
| Traumatic brain dysfunction | 20,912 (4.2%) |
| Non-traumatic spinal cord dysfunction | 21,516 (4.4%) |
| Traumatic spinal cord dysfunction | 4,570 (0.9%) |
| Progressive neurological conditions | 13,081 (2.7%) |
| Other neurological conditions | 56,170 (11.4%) |
| Fractures and other multiple trauma | 57,879 (11.7%) |
| Amputation | 14,622 (3.0%) |
| Other orthopedic conditions | 39,177 (7.9%) |
| Debility, cardiorespiratory conditions | 85,808 (17.4%) |
| Medically complex conditions | 7,494 (1.5%) |
| Admitted from Location |  |
| Short-term General Hospital | 458,871 (93.0%) |
| Home (with or without home care) | 19,378 (3.9%) |
| Post-Acute Care\*\*\*\* | 11,517 (2.3%) |
| Other† | 2,937 (0.6%) |
| Not Listed | 506 (0.1%) |
| Discharge to Location |  |
| Short-Term General Hospital | 49,206 (10.0%) |
| Home (with or without home care) | 364,486 (73.9%) |
| Post-Acute Care\*\*\*\* | 74,379 (15.1%) |
| Other† | 4,038 (0.8%) |
| Not Listed | 1,100 (0.2%) |

Note: Values are reported as frequency (percent)

\*Percentages can add up to more than 100%; if more than 1 category was selected the patient is assigned to both categories.

\*\*Includes divorced, separated, never married, and not assessed/no information.

\*\*\*Includes friend, attendant, other person, and not assessed/no information.

\*\*\*\*Includes institutional settings: skilled nursing facilities, long-term care hospitals, and another IRF.

† Includes nursing homes, swing beds, critical access hospitals, hospice, inpatient psychiatric facilities, and other intermediate care settings.

Source: RTI analysis of IRF-PAI, January – December 2017 (Program reference: LP57).

**Rasch Analysis Sample using National IRF-PAI Data – Fiscal Year 2017 Data**

As noted above, the reliability and validity testing that involved Rasch analysis and internal consistency testing was conducted using fiscal year 2017 data. IRF-PAI data for 163,344 randomly selected IRF patients discharged in fiscal year 2017 were analyzed for the fit assessment and internal consistency. More than half of the IRF patients were female (53.3%) and 52.3% were 75 years old or older. Most were white (79.3%) and admitted to the IRF directly from an acute care hospital (93.2%).

**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**.

Most testing was conducted using national IRF-PAI data submitted by IRFs for all Medicare Part A and Medicare Advantage patients discharged in calendar year 2017 (**Tables 1** and **2**).

For the Rasch analyses and internal consistency analyses, we used a random subsample of the national data (n = 163,344) for patients discharged in fiscal year 2017. The Rasch analysis and internal consistency work include:

* Scale Construct Validity Testing - Item Difficulty Ordering
* Scale Validity Testing - Fit Assessment and Internal Consistency
* Item Validity Testing - Response Option Assessment

**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.

We examined whether 5 social risk factors affected the computed performance measure scores: 1) dual eligibility (patient-level variable); 2) race/ethnicity (patient-level variable); 3) living alone (patient-level variable); 4) urbanicity based on the patient’s residence (community-level variable); and 5) socioeconomic status (SES) (community-level variable).

We selected the patient-level social risk factors based on our review of the literature showing functional outcomes can vary by race/ethnicity and by living situation. The selected community-level factors have been examined for other measures, but have been not addressed in the functional outcomes literature and thus the possible role and these factors have been unclear.

Dual eligibility data were derived from the Integrated Data Repository (IDR). We obtained race/ethnicity and living alone status from the IRF-PAI. Urbanicity was defined by cross-walking beneficiary residence ZIP codes (from the IRF-PAI) to Federal Information Processing Standard Publication (FIPS) codes,[[1]](#footnote-1) then cross-walking FIPS codes to Rural-Urban Commuting Area Codes (RUCA\_2013).[[2]](#footnote-2) Socioeconomic status was determined using the Agency of Healthcare Research and Quality’s SES Index[[3]](#footnote-3) calculated based on the patient’s residence ZIP Code Tabulation Area (ZCTA). ZCTA was found by cross-walking the beneficiary residence ZIP code with ZCTA. We used data from the 2016 American Community Survey (5-year file) to calculate AHRQ SES Index, with higher values indicating higher SES.

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

**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*)

We report testing results throughout this document for data elements, the self-care scale/instrument and the computed performance measure score. To assist the reader in understanding the testing analysis and results, we begin by providing a brief overview of these components of the performance measure:

1. **Self-Care Data Elements:**
   1. There are 7 self-care data elements, which are included in IRF-PAI Section GG. Depending on the context, we sometimes refer to these data elements as “items” or “activities.”
   2. The self-care data are collected at the time of admission and discharge using a 6-level rating scale (01 to 06), or activity not attempted codes if, for example, the activity was not attempted due to medical or safety concerns.
   3. Higher scores indicate higher ability (i.e., more independence)
   4. For the performance measure calculation, data element activity not attempted codes and missing data are recoded to 01.
2. **Admission and Discharge Self-Care Scores (Scale/Instrument)**
   1. An admission self-care scale score is created by summing the 7 self-care data element scores, after re-coding. The admission self-care score can range from 7 to 42 self-care units.
   2. A discharge self-care scale score is created by summing the 7 data element scores, after re-coding. The range of the discharge self-care score is 7 to 42 self-care units.
   3. For the Admission and Discharge Self-Care Scores, a score of 7 indicates the patient is dependent on a helper to perform all 7 self-care activities (i.e., data elements) and a score of 42 means the patient is independent on all 7 activities.
3. **Observed Change in Self-Care** 
   1. An observed change in Self-Care score is calculated by subtracting the observed (unadjusted) Discharge Self-Care Score from the observed (unadjusted) Admission Self-Care Score.
   2. The potential range of the Observed Change in Self-Care Scores is -35 to + 35. Most patients are expected to have improved self-care abilities, and thus we observe mostly positive values.
4. **Calculated Performance Measure Score: Risk-Adjusted Change in Self-Care Score**
   1. The calculated performance measure score is a risk-adjusted Change in Self-Care Score. The risk-adjustment procedure is described in S.14. Calculation Algorithm/Measure Logic on the NQF Intent to Submit form and the attached file “IRF\_Detailed\_Function\_QM\_Specifications\_2633\_01-07-2019.docx”.
   2. This performance measure estimates the risk-adjusted mean change in self-care score between admission and discharge for IRF patients. This performance measure does not have a simple form for the numerator and denominator.

**Computed Performance Measure Score Reliability – Split-half Reliability (unit of analysis is providers):** Split-half reliability was used to examine the reliability of the computed performance measure scores. The computed performance measure scores are the risk-adjusted change in self-care scores. For IRFs with fewer than 20 patient stays, computed performance measures are not displayed to the public, therefore, we included facilities with 20 or more stays in this analysis. We conducted split-half reliability by randomly splitting each provider’s patient stays into two groups and calculating correlations between the computed performance measure scores of the randomly divided groups. When a provider’s data, after being randomly divided into two groups, show similar scores to one another, the performance measure score is more likely to reflect systematic differences in IRF provider quality rather than random variation. The Pearson Product-Moment Correlation (*r*), Spearman Rank Correlation (*ρ)*, and Intraclass Correlation Coefficient (ICC) were used to measure internal reliability. Intraclass correlations were also calculated by facility volume quartile to examine whether there were differences in performance measure reliability by IRF size.

**Self-Care Scale/Instrument Analysis - Internal Consistency (unit of analysis is patient assessments):** In addition to the provider-level reliability testing of the computed performance measure scores described above, we examined the internal consistency of the self-care scale/instrument scores for each patient-stay. Internal consistency provides a general assessment of how well the self-care data elements interrelate within the self-care scale/instrument. This internal consistency analysis is an indicator of the reliability of the self-care scale/instrument and is thus a test of the reliability of the data elements.

Internal consistency was 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 calculated when testing instrument or scale psychometrics. 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. Nunnally (1978) indicated that Cronbach’s alpha should be at least 0.90 for item sets used in decision making*.* The internal consistency from the Rasch analysis assesses items using the KR20 (a special case of Cronbach’s alpha) estimate, with the same cut-off requirements

**Citation:** Nunnally, J. (1978). *Psychometric methods*. New York, NY: McGraw-Hill.

**Critical Data Elements Testing using CARE Tool Data (2014)** **– Inter-Rater Reliability, Video (Standardized Patient) Reliability and Validity Testing (unit of analysis is patients)**: In our 2014 NQF testing document, we described several types of data element and scale/instrument reliability and validity analysis using data collected by providers as part of the Post-Acute Care Payment Reform Demonstration (2007-2012). This reliability and validity testing included the self-care and mobility data elements, as well as data elements that are used as risk adjustors for this performance measure. For more information about the development and testing of the data elements and scale/instrument, please see:

* Gage BJ, Constantine R, Aggarwal MM, Bernard S, Munevar D, Garrity M, Deutsch A, et al. (June, 2012). *The Development of the Continuity Assessment Record and Evaluation (CARE) Tool: Final Report*. Prepared for the Centers for Medicare & Medicaid Services. Available at: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/The-Development-and-Testing-of-the-Continuity-Assessment-Record-and-Evaluation-CARE-Item-Set-Final-Report-on-the-Development-of-the-CARE-Item-Set-Volume-1-of-3.pdf>
* Gage BJ, Smith LM, Ross J, Coots LA, Shamsuddin KM, Deutsch A, Mallinson T, Reilly KE, Abbate JH, Gage-Croll Z. (August, 2012). *The development and testing of the Continuity Assessment Record and Evaluation (CARE) Item Set: Final Report on Reliability Testing, Volume 2 of 3*. Prepared for Centers for Medicare & Medicaid Services. Available at: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/The-Development-and-Testing-of-the-Continuity-Assessment-Record-and-Evaluation-CARE-Item-Set-Final-Report-on-Reliability-Testing-Volume-2-of-3.pdf>
* Gage BJ, Deutsch A, Smith LM, Schwartz C, Ross J, Coots LA, Reilly KE, Abbate JH, Shamsuddin KM, Silver BC, et al. (September, 2012). *The Development and Testing of the Continuity Assessment Record and Evaluation (CARE) Item Set: Final Report on CARE Item Set and Current Assessment Comparisons, Volume 3 of 3*. Prepared for Centers for Medicare & Medicaid Services. Available at: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/The-Development-and-Testing-of-the-Continuity-Assessment-Record-and-Evaluation-CARE-Item-Set-Final-Report-on-the-Development-of-the-CARE-Item-Set-and-Current-Assessment-Comparisons-Volume-3-of-3.pdf>
* Smith LM, Deutsch A, Hand LB, Etlinger AL, Ross J, Abbate JH, Gage-Croll Z, Barch D, Gage BJ. (September, 2012). *Continuity Assessment Record and Evaluation (CARE) Item Set: Additional Provider-Type Specific Interrater Reliability Analyses*. Prepared for Centers for Medicare & Medicaid Services. Available at: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/Continuity-Assessment-Record-and-Evaluation-CARE-Item-Set-Additional-Provider-Type-Specific-Interrater-Reliability-Analyses.pdf>
* Smith LM, Deutsch A, Barch D, Ross J, Shamsuddin KM, Abbate JH, Schwartz C, Gage BJ. (September, 2012). *Continuity Assessment Record and Evaluation (CARE) Item Set: Video Reliability Testing*. Prepared for Centers for Medicare & Medicaid Services. Available at: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Downloads/Continuity-Assessment-Record-and-Evaluation-CARE-Item-Set-Video-Reliability-Testing.pdf>
* Gage BJ, Morley MA, Smith LM, Ingber MJ, Deutsch A, Kline TL, Dever JA, Abbate JH, Miller RD, Lyda-McDonald B, Kelleher CA, Garfinkel DB, Manning JR, Murtaugh CM, Stineman MG, Mallinson T. (March, 2012). *Post-Acute Care Payment Reform Demonstration: Final Report Volumes 1-4*. Prepared for the Centers for Medicare and Medicaid Services. Available at: <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Reports/Research-Reports-Items/PAC_Payment_Reform_Demo_Final.html>

For more information on the history of the development of this functional status performance measure, please visit CMS’s Post-Acute Care Quality Initiatives Function Measures website: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/Functional-Measures-.html>

**Summary of critical data element reliability testing:**

The inter-rater reliability of the data elements was tested in a subset of 34 providers (acute hospitals, HHAs, IRFs, LTCHs, and SNFs) distributed across 11 geographic areas. Each provider completed a duplicate admission or discharge assessment on 10–20 patients. The overall sample size was 450 for self-care items. 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. Unweighted overall kappas ranged from 0.636 (toileting) to 0.598 (oral hygiene). In summary, kappa statistics indicated substantial agreement of data element codes among raters.

For the video reliability study, clinicians assessed “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 scoring among different clinicians examining the “same” patient. 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 LPNs); they consistently had the lowest levels of agreement among 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. Because the clinical team response and mode were identical on most of the videos, agreement was still quite high for these items.

Please see Appendix B for the additional details about the inter-rater reliability and video reliability testing.

**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*)

**Computed Performance Measure Score Reliability (unit of analysis is provider):** Split-half analysis results indicated strong, positive correlations (r = 0.903, ρ = 0.884, ICC= 0.903, p < 0.0001) between the IRF providers’ randomly divided groups’ computed performance measure scores for the Change in Self-Care performance measure, providing strong evidence of measure reliability. As shown in **Table 3**, ICCs remained strong when stratifying by provider volume quartile, with ICCs for the volume quartiles ranging from 0.836 (20-174 discharges) to 0.965 (568 - 4,416 discharges).

Table 3. Interclass Correlation Coefficient by IRF Volume, Calendar Year 2017 (N=1,117)

| **Volume Quartile** | **Number of IRFs** | **ICC** |
| --- | --- | --- |
| **Quartile 1: 20 - 174** | 280 | 0.836 |
| **Quartile 2: 175 - 295** | 278 | 0.899 |
| **Quartile 3: 296 - 566** | 280 | 0.944 |
| **Quartile 4: 568 - 4,416** | 279 | 0.965 |
| **Total** | 1,117 | 0.903 |

Note: Providers with < 20 stays during the 12-month testing period are excluded.

Source: RTI analysis of IRF-PAI January – December 2017 (Program reference: MV52).

**Scale/Instrument Reliability - Internal Consistency (unit of analysis is patient stays):** Analysis of the self-care data showed good reliability statistics. The overall Cronbach’s alpha is 0.94.

**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 analysis of calendar year 2017 data show that provider-level reliability of the computed performance measure scores was strong overall and when stratified by provider volume. The patient-level analysis of fiscal year 2017 data of the scale/instrument reliability showed very good reliability.

Critical data element inter-rater reliability and video reliability testing found substantial reliability overall.

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

**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)*

**Scale/Instrument Content Validity - Similarity of Data Elements Across Other Self-Care Assessment Instruments:** Patient functioning is a construct that is often measured based on patient abilities, and the activities (data elements) included in functional assessment instruments vary. We compared the list of Section GG data elements used to calculate the Change in Self-Care performance measure with self-care data elements included on other functional assessment instruments.

**Face Validity – Technical Expert Survey** - On March 27, 2017, RTI International, on behalf of the Centers for Medicare & Medicaid Services (CMS), convened an in-person Technical Expert Panel (TEP) in Baltimore, MD, to seek expert input on the Inpatient Rehabilitation Facilities Quality Reporting Program (IRF QRP) quality measures, including the functional status performance measures. Prior to the TEP meeting, TEP members provided feedback on the importance, scientific soundness and usability of each of the performance measures using a 5-level Likert scale (high, moderately high, neutral, moderately low, low).

**Data Element Construct Validity – Observed Discharge Self-Care Scores and Discharge Destination (unit of analysis is patient stays):** We tested the validity of the self-care data by examining the discharge function scores and whether patients were discharged to a community destination. IRFs provide intensive rehabilitation services to patients with a goal of maximizing patient functioning so that the patient can be ideally discharged home and avoid institutionalization. IRF patients who have higher abilities are more likely to be discharged to their home or another community-based setting compared to patients discharged to another post-acute care setting (e.g., skilled nursing facility, long-term care hospital), nursing home, hospice, or an acute-care hospital. Therefore, we tested the construct validity of the self-care data by examining the relation between discharge self-care scores and being discharged to the community, after excluding incomplete stays.

**Scale/Instrument Construct Validity – Observed Discharge Self-Care Scores and Discharge Destination (unit of analysis is patient stays):** We tested the validity of the scale/instrument scores by examining the discharge self-care scale scores and whether patients were discharged to a community destination. We ran a logistic regression model to examine the association between discharge self-care scores and the odds of a community discharge.

**Scale/Instrument Construct Validity - Data Element (Item) Difficulty Ordering Using Rasch Analysis (unit of analysis is patient assessment data):** Rasch analysis uses item data to determine how well items in a scale/instrument function together to measure a construct. In its base form, the Rasch model assumes that the probability of a code for a given item is a function of the patient’s self-care ability and the item’s difficulty (how hard the activity is to accomplish independently). The Rasch extension that accounts for multiple response options also considers the difficulty of moving from one code category to another (i.e., a threshold). The information resulting from this function is interval in nature and expressed on the log-odds scale. Also, as part of the analysis, Rasch methodology places persons and the items of interest on a “ruler” to enable evaluations of how well the items work together, how difficult each item is relative to the other items in the scale/instrument, and how items are ordered from easy to difficult. We used Rasch measurement analysis to examine the self-care items. We report IRF analysis results using a Rasch-derived self-care ruler that was developed using data from IRFs, skilled nursing facilities and long-term care hospitals. Using the Rasch-derived cross-setting “ruler” allows comparability of self-care item functioning within and across settings.

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) and against expert clinical opinions as an indication of construct validity. If items are positioned into unexpected locations on the hierarchy, then the content of the items should be evaluated further and potentially modified.

**Data Element (Item) and Scale/Instrument Validity - Fit Assessment Analysis (unit of analysis is patient assessment data):** Rasch analysis produces fit statistics that reflect whether unexpected responses are being coded for items within the scale/instrument. The Rasch model expects the difficult items to be harder (that is, have a greater need for assistance) for all patients. In a similar way, patients with higher functional abilities are generally expected to need less assistance on all items. Items that don’t seem to function this way could show misfit, reflecting unexpected responses. There are two categories of fit, one designed more for outliers (outfit) and one designed for response unexpectedness near the item’s difficulty (infit). In general, a cut-off appropriate for statistically determining item misfit is infit and outfit mean square values are above 1.4 when looking at multiple-point response scales. Items with fit values above 1.4 are unproductive for measurement but are not unusually “noisy” or degrade measurement. Mean square values greater than 2.0 may potentially degrade measurement (Wright and Linacre, 1994). 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.

**Data Element (Item) and Scale/Instrument Validity - Response Option Assessment Using Rasch Analysis (unit of analysis is patient assessment data):** Rasch analysis output reports the number and percent of patients by score level (06 - Independent to 01 - Dependent) for each item and the average self-care ability (i.e., scale-level ability) of those patients. This allows us to examine if the 6-point rating scale is operating as intended for the self-care items. In general, we expect that patients who have lower ability overall would have lower ability levels (i.e., lower scores) for each item. Therefore, the average self-care ability calibration (scale-level ability measure reported in logits) associated with the more dependent scores would be lower than those associated with the more independent scores.

**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>

**Computed Performance Measure Score Validity – Association with The Joint Commission Stroke Rehabilitation Certification Status (unit of analysis is providers):** The goal of measuring performance is to make valid (credible) conclusions about quality (NQF Committee Guidebook). To examine the validity of the Change in Self-Care computed performance measure score, we conducted analyses using a structural measure of quality, whether or not an IRF obtained The Joint Commission’s Disease Specific Certification for Stroke Rehabilitation. As previously noted in Table 1, stroke is the most common primary medical condition for patients admitted to IRFs, therefore stroke patient outcomes influence IRF performance measure scores. The Joint Commission’s Disease-Specific Care Certification evaluates clinical programs addressing: 1) Compliance with consensus-based national standards; 2) Effective use of evidence-based clinical practice guidelines to manage and optimize care; and 3) An organized approach to performance measurement and improvement activities. According to The Joint Commission, an entity that achieves Disease-Specific Certification has thoroughly demonstrated a high level of care for patients with that condition. We downloaded data from The Joint Commission’s website and we used an ‘effective date’ to identify IRFs that were certified during the calendar year 2017. More information about disease-specific certification, please see: <https://www.jointcommission.org/certification/dsc_physical_medicine_rehabilitation.aspx>

Our first analysis compared the mean and median computed performance measure scores for IRFs with and without stroke rehabilitation disease-specific certification using a t-test and Kruskal-Wallis H test. We expected that IRFs with certification would achieve higher mean and median performance measure scores compared to IRFs without certification. Second, we divided the IRF data into quintiles based on the performance measure scores and calculated the percentage of IRFs with certification by quintile. We expected that IRFs with the best performance scores (quintile 5) would have a higher percentage of certified IRFs compared to the IRFs in quintile 1 with the least favorable performance measure scores.

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

**Content Validity: Similarity of Data Elements Across Other Self-Care Assessment Instruments.**

Many functional status scales have been developed for research and clinical use. To address content validity, we have provided a table listing activities (data elements) used to calculate the Change in Self-Care performance measure and data elements included in other functional assessment scales. **Table 4**, shows that the Section GG self-care activities cover a wide range of self-care activities and that many of the activities included on other instruments (e.g., Eating, Toileting hygiene) are included in Section GG.

**Table 4. Comparison of Selected Self-Care Activities (Data Elements) for the Change in Self-Care Performance Measure and Other Functional Assessment Instruments.**

| **Activity (Data Elements)** | **Change in Self-Care Measure: Section GG Self-Care Items** | **Barthel Index** | **FIM® Instrument** | **Katz ADL Scale** | **FAM with FIM®**  **Instrument** | **RICFAS with FIM® Instrument** | **Minimum Data Set: Section G** | **Upper Extremity Functional Index** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Eating**  **ICF =** Drink d560  Eat d550 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |  |
| **Grooming/personal or oral hygiene**  **ICF =** Caring for teeth d5201 | ✓ | ✓ | ✓ |  | ✓ | ✓ | ✓ | ✓ |
| **Toileting hygiene**  **ICF =** Regulating Urination d5300  Regulating Defecation d5301 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |  |
| **Bathing/showering**  **ICF = d**5100 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |  |
| **Dressing: lower body**  **ICF =** Putting on clothes d5400  Taking off clothes d5401 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Dressing: upper body**  **ICF =** Putting on clothes d5400  Taking off clothes d5401 | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Dressing: footwear**  **ICF =** Putting on clothes d5400  Taking off clothes d5401 | ✓ |  |  |  | ✓ |

Notes: ADL = activity of daily living; ICF = International Classification of Functioning; FAM = Functional Assessment Measure; RICFAS = Rehabilitation Institute of Chicago Functional Assessment Scale

**F****ace Validity – Technical Expert Survey:** For the Change in Self-Care performance measure, 71% of TEP members rated the Measure Importance as High or Moderately High; 57% rated the Scientific Soundness as High or Moderately High, and 57% Rated Usability of the Measure as High. We note that this survey was conducted prior to the TEP meeting, and thus represents perceptions before the TEP discussions about the measure details and measure testing results. In addition, this TEP occurred approximately 8 months after the implementation of data collection when confidential feedback reports were not yet available to providers. Finally, with the goal of learning from experts in order to drive measure improvement efforts, for this TEP we invited representatives from organizations that had previously given feedback on the measures and that had competing measures. Thus, full support for the measure was not an expected outcome of the pre-TEP survey, and the survey provided TEP members an opportunity to give constructive feedback based on their initial perceptions before participating in the panel.

**Data Element Construct Validity: Observed Discharge Functional Ability and Discharge Destination (unit of analysis is patient stays):** As shown in **Table 5,** patients with higher discharge scores (from 01 – Dependent to 06 – Independent), are more likely to be discharged to the community, as expected. This occurs for each self-care data element for all score levels, with the exception of the data element Eating level 1, which has a slightly higher percentage compared to level 2. Also expected, for each of the self-care data elements (**Table 5**), patients who were coded as 06 - Independent, a high percentage were discharged to the community (89.6% for Eating to 97.3% for Shower/Bathe Self).

Findings and Interpretation: Self-care data were positively associated with discharge destination, as expected. Specifically, we found patients who had higher observed scores at discharge were more likely to be discharged to a community setting, which supports the validity of the item data measuring functional abilities in the IRF population.

**Table 5. Observed Discharge Self-Care Data Element Scores and Discharge Location (n=437,619)**

| **Data Element/Score** | **Discharged to Community** | |
| --- | --- | --- |
| **GG0130A3: Self-Care - Eating** |  |  |
| 01-Dependent | 2,131 (43.8%) |  |
| 02-Substantial/maximal assistance | 1,114 (40.7%) |  |
| 03-Partial/moderate assistance | 2,294 (48.7%) |  |
| 04-Supervision or touching assistance | 15,109 (61.5%) |  |
| 05-Setup or clean-up assistance | 55,914 (68.0%) |  |
| 06-Independent | 281,212 (89.6%) |  |
| **GG0130B3: Self-Care - Oral Hygiene** |  |  |
| 01-Dependent | 1,227 (34.6%) |  |
| 02-Substantial/maximal assistance | 1,508 (38.8%) |  |
| 03-Partial/moderate assistance | 4,440 (47.7%) |  |
| 04-Supervision or touching assistance | 28,160 (66.3%) |  |
| 05-Setup or clean-up assistance | 62,451 (68.5%) |  |
| 06-Independent | 260,236 (91.8%) |  |
| **GG0130C3: Self-Care - Toileting Hygiene** |  |  |
| 01-Dependent | 11,071 (38.7%) |  |
| 02-Substantial/maximal assistance | 14,430 (52.8%) |  |
| 03-Partial/moderate assistance | 25,697 (67.2%) |  |
| 04-Supervision or touching assistance | 72,684 (79.8%) |  |
| 05-Setup or clean-up assistance | 31,936 (84.0%) |  |
| 06-Independent | 202,239 (96.2%) |  |
| **GG0130E3: Self-Care - Shower/Bathe Self** |  |  |
| 01-Dependent | 3,155 (34.5%) |  |
| 02-Substantial/maximal assistance | 8,215 (44.1%) |  |
| 03-Partial/moderate assistance | 45,471 (63.6%) |  |
| 04-Supervision or touching assistance | 97,907 (82.5%) |  |
| 05-Setup or clean-up assistance | 55,335 (88.3%) |  |
| 06-Independent | 145,718 (97.3%) |  |
| **GG0130F3: Self-Care - Upper Body Dressing** |  |  |
| 01-Dependent | 2,124 (35.5%) |  |
| 02-Substantial/maximal assistance | 5,792 (43.4%) |  |
| 03-Partial/moderate assistance | 24,826 (60.2%) |  |
| 04-Supervision or touching assistance | 43,858 (73.2%) |  |
| 05-Setup or clean-up assistance | 67,130 (76.4%) |  |
| 06-Independent | 215,723 (94.9%) |  |
| **GG0130G3: Self-Care - Lower Body Dressing** |  |  |
| 01-Dependent | 8,562 (38.5%) |  |
| 02-Substantial/maximal assistance | 17,755 (52.5%) |  |
| 03-Partial/moderate assistance | 39,276 (69.1%) |  |
| 04-Supervision or touching assistance | 75,410 (81.2%) |  |
| 05-Setup or clean-up assistance | 41,270 (87.0%) |  |
| 06-Independent | 177,127 (97.2%) |  |
| **GG0130H3: Self-Care - Putting on/Taking Off Footwear** |  |  |
| 01-Dependent | 16,689 (46.5%) |  |
| 02-Substantial/maximal assistance | 18,191 (57.3%) |  |
| 03-Partial/moderate assistance | 34,789 (72.6%) |  |
| 04-Supervision or touching assistance | 57,825 (80.7%) |  |
| 05-Setup or clean-up assistance | 48,402 (84.3%) |  |
| 06-Independent | 180,806 (96.6%) |  |

Notes: Values reported as frequency (percent); Incomplete stays are excluded; Activity not attempted codes not shown.

Source: RTI analysis of IRF-PAI, January– December 2017. (Program reference: LP63).

**Scale/Instrument Construct Validity: Observed Discharge Functional Ability and Discharge Destination (unit of analysis is patient stays). Table 6** displays the single variable logistic regression results with observed discharge self-care scale scores as the independent variable and a dichotomous dependent variable indicating whether the IRF patient was discharged to the community or not. The self-care scale score is the sum of the 7 self-care data element scores after recoding; the discharge self-care scale scores can range from 7 to 42. The results show that, on average, a one-unit increase in discharge self-care score is associated with a 16 percent increase in the odds of being discharged to the community (OR = 1.159; *p*-value <0.001).

Findings and Interpretation: Self-care scale/instrument scores were positively associated with discharge destination, as expected. Specifically, we found patients who had higher observed scores at discharge were more likely to be discharged to a community setting, which supports the validity of the scale/instrument data measuring functional abilities in the IRF population.

Table 6. Coefficient and Odds Ratio for Discharge to Community Model (n=437,619)

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | **Value** | **95% Confidence Interval** |
| Observed Discharge Self-Care Score |  |  |
| Coefficient | 0.148 |  |
| Odds Ratio | 1.159 | 1.158 – 1.161 |

Note: Observed discharge self-care score range = 7 – 42; Incomplete stays were excluded.

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: LP63).

**Scale/Instrument Construct Validity: Data Element (Item) Difficulty Ordering Using Rasch Analysis (unit of analysis is patient assessment data):** We usedRasch analysis to determine how well the self-care items work together to measure the construct of self-care. Rasch analysis creates a self-care ruler using log odd units (i.e., logits) centered at the value 0. A “logit” (a contraction of "Log-Odds Unit") is a linear scale. We report IRF testing results using a Rasch-derived self-care ruler that was developed using data from IRFs, skilled nursing facilities and long-term care hospitals. The analysis of the Section GG self-care data show that the placement of each self-care item on the cross-setting self-care “ruler” make sense clinically and are consistent with previous analyses of other functional assessment scales. That is, the order of items from easy to difficult (item hierarchy), is consistent with task difficulties. The order of the items by difficulty level, with the hardest activity listed first, is as follows:

Footwear (most difficult activity)

Lower Body Dressing

Shower

Toileting Hygiene

Upper Body Dressing

Oral Hygiene

Eating (easiest activity)

**Figure 1** reports the item hierarchy, the patient distribution and the rating scale scores in one graphic. In addition, **Figure 1** is presented on theRasch-derived self-care ruler, expressed in logits and centered at a value of 0, as described previously. It shows the overall expected score placement on the self-care “ruler” for each item. The ruler values, ranging from -7 to +7 logits, are shown on the top and bottom vertical lines. The difficulty order (item hierarchy), 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 being 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.”

**Findings and Interpretation:** The item hierarchy listing and **Figure 1** illustrate that the self-care items fall along the cross-setting “ruler” as expected and are consistent with clinical findings from applications in the field and other functional assessment instruments.

**Figure 1. Self-Care IRF Items – Anchored on the Cross-Setting Self-Care Ruler**

-7 -5 -3 -1 1 3 5 7

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

1 1 : 2 : 3 : 4 : 5 : 6 6 19\* Footwear

1 1 : 2 : 3 : 4 : 5 : 6 6 18\* Lower Body Dressing

1 1 : 2 : 3 : 4 : 5 : 6 6 16\* Shower

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

| |

1 1 : 2 : 3 : 4 : 5 : 6 6 17\* Upper Body Dressing

| |

| |

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

| |

1 1 : 2 : 3 : 4 : 5 : 6 6 1\* Eating

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

-7 -5 -3 -1 1 3 5 7

T S M S

0 10 20 30 50 60 70 80 99 PERCENTILE

**Scale/Instrument Validity - Fit Assessment Using Rasch Analysis (unit of analysis is patient assessment data):** Ideal measurement construction would mean data fit the Rasch model exactly. In reality, empirical data will differ from the model. Rasch fit statistics describe how well the observed data (e.g. patient’s scores on the self-care items) fit the model, and characterize the magnitude that unexpected scores (i.e., unmodelled noise) are found in the data. Fit statistics have an expected value of 1.0 and can range from 0 to infinity. Values lower than 1.0 indicate overfit (over prediction) of the Rasch model and values greater than 1.0 indicate underfit of the model (e.g., noise). There are two categories of fit. Outfit is designed more for outliers (when a patient’s unexpected code is for an item that is relatively easy or hard for that patient); Infit is designed for unexpected codes near the item’s difficulty (when a patient’s code is for an item is near that person’s ability). Values greater than 2.0 may potentially degrade measurement (Wright and Linacre, 1994). Overall, the self-care items are coded as expected. **Table 7** reports fit statistics for the self-care items and shows that one item, Eating, had fit statistics above 2.00. The item Eating has been found to misfit in other functional assessment scales, and this misfit may be related to patients who have a swallowing impairment.

**Table 7. Fit Statistics for the Self-Care Items (n = 321,392)**

|  |  |  |
| --- | --- | --- |
|  | **IRF – Anchored**  **(Cross-Setting Ruler)** | |
| Item | Infit mean square | Outfit mean square |
| GG0130A: Eating | 1.56 | 2.38 |
| GG0130B: Oral Hygiene | 1.14 | 1.16 |
| GG0130C: Toileting Hygiene | 1.15 | 1.15 |
| GG0130E: Shower/Bathe Self | 0.79 | 1.04 |
| GG01030F: Upper Body Dressing | 0.98 | 0.99 |
| GG0130G: Lower Body Dressing | 0.64 | 0.67 |
| GG0130H: Putting On/Taking Off Footwear | 1.19 | 1.14 |

**Data Element (Item) and Scale/Instrument Validity - Response Option Assessment Based on Rasch Analysis (unit of analysis is patient assessments):** Rasch analyses provide information on how many patients are coded in each score category (i.e., independent to dependent) for each item and the average ability (or skill level) of those individuals on the construct of interest. Evaluations of patient ability by score category indicate that rating scale use is as expected, with patients with higher item scores are, on average, higher ability patients. For our data, we anticipate that for each item, patients with higher scores (01 to 06) should have higher Rasch logit self-care values (Rasch self-care logit values range from -7 to +7). Likewise, 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 scores would be lower than ability estimates associated with the more independent scores. We combined admission and discharge data for each item in order to ensure a range of patient ability is represented in the analyses.

As shown in **Table 8**, for each item, patients who are coded with higher scores have higher overall self-care ability, as expected.

**Table 8. Distribution of Combined Admission and Discharge Scores and Average Ability Estimate by Response Code for Each Self-Care Item (n = 321,392)**

| **Item** | **Score (Response Code)\***  **Higher Score = Higher Ability** | **Number of Patients** | **Percent of Patients by Item** | **Average Self-Care Ability of Patients**  **(- 7 to +7 Logit Scale) Higher Value = Higher Ability)** |
| --- | --- | --- | --- | --- |
| Eating |  |  |  |  |
|  | 01 | 7503 | 2 | -4.33 |
|  | 02 | 4778 | 2 | -3.21 |
|  | 03 | 10118 | 3 | -2.09 |
|  | 04 | 32476 | 10 | -0.83 |
|  | 05 | 103346 | 33 | -0.03 |
|  | 06 | 156927 | 50 | 3.16 |
| Oral hygiene | |  |  |  |
|  | 01 | 7490 | 2 | -5.04 |
|  | 02 | 7832 | 3 | -2.84 |
|  | 03 | 20976 | 7 | -1.61 |
|  | 04 | 58049 | 19 | -0.30 |
|  | 05 | 112306 | 36 | 0.30 |
|  | 06 | 106575 | 34 | 4.44 |
| Toileting hygiene | |  |  |  |
|  | 01 | 50601 | 16 | -2.18 |
|  | 02 | 44380 | 14 | -0.84 |
|  | 03 | 49142 | 16 | 0.08 |
|  | 04 | 68799 | 22 | 1.13 |
|  | 05 | 21247 | 7 | 2.18 |
|  | 06 | 75748 | 24 | 5.52 |
| Shower/bathe self | |  |  |  |
|  | 01 | 19958 | 7 | -3.24 |
|  | 02 | 39856 | 13 | -1.42 |
|  | 03 | 95554 | 32 | -0.09 |
|  | 04 | 69272 | 23 | 1.60 |
|  | 05 | 27569 | 9 | 3.22 |
|  | 06 | 50838 | 17 | 6.47 |
| Upper body dressing | |  |  |  |
|  | 01 | 15196 | 5 | -3.81 |
|  | 02 | 28103 | 9 | -1.82 |
|  | 03 | 62352 | 20 | -0.63 |
|  | 04 | 57758 | 18 | 0.40 |
|  | 05 | 73505 | 23 | 1.18 |
|  | 06 | 79272 | 25 | 5.46 |
| Lower body dressing | |  |  |  |
|  | 01 | 53550 | 17 | -2.22 |
|  | 02 | 63504 | 20 | -0.71 |
|  | 03 | 59284 | 19 | 0.39 |
|  | 04 | 56764 | 18 | 1.51 |
|  | 05 | 21514 | 7 | 2.86 |
|  | 06 | 61990 | 20 | 6.26 |
| Putting on/taking off footwear | | | |  |
|  | 01 | 78309 | 25 | -1.65 |
|  | 02 | 51625 | 17 | -0.53 |
|  | 03 | 42052 | 13 | 0.55 |
|  | 04 | 44284 | 14 | 1.37 |
|  | 05 | 31622 | 10 | 2.35 |
|  | 06 | 64390 | 21 | 6.09 |

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.

**Computed Performance Measure Score Validity – Association with The Joint Commission Stroke Rehabilitation Certification Status (unit of analysis is providers):** We compared the mean and median computed performance measure scores for IRFs with and without stroke rehabilitation disease-specific certification. We also divided the IRF data into quintiles based on the performance measure scores and calculated the percentage of IRFs with certification by quintile.

**Table 9** shows that IRFs with certification achieved slightly higher mean and median performance measure scores compared to IRFs without certification (mean: 12.1 and 11.4 and p < .0001; median 12.0 and 11.3 and p < .001).

**Table 9. Mean and Median Change in Self-Care Computed Performance Measure Score (CY 2017) by Stroke Rehabilitation Disease Specific Certification Status (2017) (n = 1,117)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Change in Self-Care Performance Measure Score** | **Stroke Rehabilitation**  **Disease Specific Certification Status**  **(2017)** | |  |
|  | **No (n=941)** | **Yes (n=176)** | **p-value\*** |
| Mean (SD) | 11.4 (1.7) | 12.1 (1.4) | < 0.001 |
| Median (IQR) | 11.3 (2.1) | 12.0 (1.6) | < 0.001 |

Note: SD=Standard deviation; IQR=Interquartile range; Providers with <20 stays during the 12-month testing period are excluded from facility-level analyses.

\*T-test was run to determine statistically significant differences for the mean scores; The Kruskal-Wallis H test was run to determine statistically significant differences for the median scores.

Source: RTI analysis of IRF-PAI, January – December 2017. (Program references: AD01)

**Table 10** shows that the top 3 quintiles, which included the IRFs with the best performance scores, had the highest percentage of certified IRFs (29.0%, 25.0%, and 25.6%) compared to the lowest quintile with the lowest performance measure scores (9.1%).

**Table 10. Percent of IRF with Stroke Rehabilitation Disease Specific Certification by Computed Performance Measure Score (CY 2017) Quintiles (n = 1,117)**

|  |  |  |
| --- | --- | --- |
| **Quintile Group Based on Performance Measure Score:**  **Best to Worst** | **Stroke Rehabilitation**  **Disease Specific Certification Status**  **(2017)\*** | |
|  | **No**  **(n=941)** | **Yes**  **(n=176)** |
| Quintile 5: 12.9-17.5  (best performance scores) | 177 (18.8%) | 51 (29.0%) |
| Quintile 4: 11.9-12.8 | 167 (17.7%) | 44 (25.0%) |
| Quintile 3: 11.1-11.8 | 190 (20.2%) | 45 (25.6%) |
| Quintile 2: 10.2-11.0 | 200 (21.3%) | 20 (11.4%) |
| Quintile 1: 5.4-10.1  (worst performance scores) | 207 (22.0%) | 16 (9.1%) |

Note: Providers with <20 stays during the 12-month testing period are excluded from facility-level analyses.

\*Chi square test results: p < .0001

Source: RTI analysis of IRF-PAI, January – December 2017. (Program references: AD01)

**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 activities (data elements) included in the Section GG self-care scale/instrument cover a wide range of patient functioning and key activities included in many other functional assessment instruments, supporting content validity of the scale.

Prior to their participation in the TEP, the panel members were surveyed on their initial perceptions of the Change in Self-Care performance measure. Most experts convened indicated the performance measure was important, scientifically sound, and able to be used by providers, patients, and the general public.

We found that patients who had higher observed discharge scores for the self-care data elements were more likely to be discharged to the community, as expected. Results also showed that the observed self-care scale/instrument scores were significantly associated with being discharged to the community.

The difficulty order of the self-care data elements makes sense clinically and are consistent with previous analyses of the self-care data and analyses of other functional assessment scales/instruments. Rasch analysis of the data showed the items work well together to measure the concept of self-care, with generally good infit and outfit statistics. As expected, for each item, the average self-care ability Rasch measure of patients increases as the rating scale scores increase. All these results support the validity of the self-care data elements and scale in measuring self-care functional abilities.

Our analyses that focused on whether or not an IRF obtained The Joint Commission’s Disease Specific Certification for Stroke Rehabilitation showed that IRFs with higher (better) computed performance measure scores were more likely to have this structural measure of quality (certification). These analyses support the validity of the calculated performance measure scores.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

**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*)

We examined the number and percentage of patients who were excluded from the performance measure calculation due to exclusion criteria. The exclusion criteria are applied to the data in order to maintain the validity of the calculated performance measure scores and were identified in consultation with expert panel members and in response to public comments. Some IRFs specialize in the care of patients with complex needs, for example, patients with traumatic spinal cord injury and traumatic brain injury; therefore, application of these exclusion criteria is important to ensure the validity of the calculated performance scores for all IRFs, regardless of whether the IRF offers specialized services for complex patients. All exclusion criteria were applied prior to our developing the risk-adjustment model.

For several exclusion criteria, the rationale for the exclusion of these patients is that improvement in self-care would be limited or unpredictable. For these exclusion criteria, we report the mean, median and 25th and 75th percentiles for change in self-care scores.

For patients who have an incomplete stay (e.g., emergency discharge), it is challenging to collect accurate discharge functional status data due to the urgent nature of the discharge. Therefore, patients with incomplete stays are excluded from the performance measure calculation, and we are unable to conduct analyses due to the unavailability of data. A total of 55,590 (11.3%) of patient stays were classified as incomplete stays based on the definition of an incomplete stay.

We excluded patients younger than 21 in our original measure specifications because we had very few patients in our sample younger than 21 and there is limited literature about functional outcomes for Medicare patients younger than 21. We are maintaining this exclusion criterion because there is still limited evidence in the literature about function outcomes for Medicare beneficiaries who are younger than 21 and there were only 32 patients younger than 21 discharged in calendar year 2017.

**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*)

A total of 65,017 patient stays (13.2%) are excluded from the change in self-care performance measure. As indicated above, most of these (55,590 (11.3%)) are due to incomplete stays. An analysis of differences between patient-level characteristics for those included and excluded from the performance measure (available upon request) show very little variation in the two populations. The largest difference was 1.1% and observed for gender (53.0% and 54.1% identified as female for the full population and the population with exclusions applied, respectively). As noted above, these exclusion criteria are important to apply to ensure the validity of the calculated performance scores for all IRFs, regardless of whether the IRF offers specialized services for complex patients.

**Table 11** shows the number and percent of patients excluded for each exclusion criteria, and the mean, median and 25th and 75th percentile for the change in self-care scores (values are reported as units of change in self-care, possible range: -35 to 35). For patients with persistent vegetative state, locked-in syndrome, those discharged to hospice and patients who are independent with all self-care activities on admission, analyses show these patients had very limited improvement with the self-care activities. For patients in a coma and those with severe brain damage, severe anoxic brain damage, and cerebral edema, improvement in self-care showed variability when we examined unadjusted data by quarter.

Table 11. Observed Change in Self-Care Score in Self-Care Units by Exclusion Criteria\* (N=493,209)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Exclusion Criteria** | **n (%)** | **Mean** | **SD** | **Median** | **25th  Percentile** | **75th  Percentile** |
| All Excluded Medical Conditions | 7,650 (1.6) | 8.8 | 8.1 | 9.0 | 4.0 | 14.0 |
| Coma | 65 (<0.1) | 8.6 | 7.0 | 9.0 | 3.0 | 14.0 |
| Complete Tetraplegia | 311 (0.1) | 4.7 | 6.8 | 3.0 | 0.0 | 8.0 |
| Persistent vegetative state\*\* | < 11 (<0.1) | \*\* | \*\* | \*\* | \*\* | \*\* |
| Severe brain damage | 731 (0.1) | 9.1 | 7.7 | 9.0 | 4.0 | 14.0 |
| Locked-In Syndrome | 12 (<0.1) | 3.6 | 6.7 | 0.0 | 0.0 | 6.0 |
| Severe anoxic brain damage, cerebral edema, or compression of the brain | 6,631 (1.3) | 8.9 | 8.1 | 9.0 | 4.0 | 14.0 |
| Discharged to Hospice | 2,548 (0.5) | 2.2 | 7.8 | 1.0 | -2.0 | 7.0 |
| Independent with all Admission Self-Care Data Elements | 714 (0.1) | -2.8 | 7.1 | 0.0 | -1.0 | 0.0 |

Note: N = number of patient stays; Observed Change in Self-Care values are reported as units of change in self-care (possible range: -35 to 35)

\*For patients who have an incomplete stay (e.g., emergency discharge), it is challenging to collect accurate discharge functional status data. Therefore, we are unable to conduct analyses due to the unavailability of IRF-PAI data. For the exclusion criterion age younger than 21, we have not conducted analyses due to the very small number of patients in this age group. In calendar year 2017, there were 32 patients younger than 21.

\*\*The number of patients with this medical condition is less than 11, and thus too small to publicly report.

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV47)

**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*)

In calendar year 2017 data, 13.2% of patient stays were excluded from the calculated performance scores. The exclusion criteria are applied to the data in order to maintain the validity of the calculated performance measure scores. Data analysis results support these exclusions, because inclusion of limited and less predictable self-care improvement for these patients could affect computed performance measure scores for the selected IRFs that admit patients who meet these criteria.

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

**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*** [***2b4***](#section2b5)***.***

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

**No risk adjustment or stratification**

**Statistical risk model with** 93 **risk factors**

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

**Other,** Click here to enter description

**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 risk adjustment model, including the intercept (constant), covariates (risk factors) with definitions and coefficients are provided as an attached excel file and in Appendix A Table A-1. We used a Generalized Linear Model regression analysis to obtain the regression intercept (constant) and regression coefficients values.

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

The risk-adjustment model includes a total of 93 covariates. For each individual patient, not every covariate will apply, because, for example, only one age group, one primary diagnosis group, and one bladder incontinence covariate will apply. In addition, patients could have 0 or up to 40 comorbidities. Therefore, for an individual patient stay, up to 61 covariates may apply.

As described in the measure calculation algorithm, the regression intercept and coefficients are used to calculate an expected change in self-care score for each patient stay using the formula below:

*Expected change in self-care score =*

intercept + (age group \* coefficient) + (continuous admission self-care \* coefficient) + (squared admission self-care \* coefficient) + (primary diagnosis group \* coefficient) + (interaction term for admission self-care and primary diagnosis group \* coefficient) + (prior surgery \* 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 \* coefficient) + (swallowing ability: tube/parenteral feeding \* coefficient) + (low BMI \* 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. Patients could have multiple comorbidities.

We provide detailed measure calculation instructions for this performance measure in an attachment in the “NQF Specifications” document. The detailed measure calculation instructions are available to the public in the document entitled “IRF Quality Reporting Program Measure Calculations and Reporting User’s Manual” that can be found at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-Quality-Reporting-Program-Measures-Information-.html. The current version of the manual, Version 3.0, reflects current measure specifications. We will implement the changes described in the NQF specifications document in 2019.

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

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

1. **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.
2. **Mean expected change in self-care score for each IRF:** We calculate each patient’s expected change in self-care score using the intercept and coefficient values that apply to the patient. We then compute 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.
3. **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 is calculated using the formula:

**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**.

Not applicable. This performance measure is risk adjusted.

**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?

This performance 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 characteristics (e.g., diagnosis) present at the time of admission that may influence functional outcomes, to allow change in self-care outcomes to be compared across IRFs.

Initial development of the risk adjustment model can be found on this measure’s previous testing form. We are now updating the risk adjustment model for this measure using the national data collected using the Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-PAI) including patients’ primary conditions, prior functioning, and comorbidities at admission. Testing of the risk adjustment model was conducted after applying the exclusion criteria described in 2b2.

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

The initial selection of risk adjustors was based on a review of the literature, input from technical experts and public comments, followed by data analysis. Please see the 2014 testing form on this measure for more detailed information on the initial selection of risk adjustors for this measure. In preparation for endorsement maintenance, we updated our literature review and conducted additional analyses.

We tested the risk adjustors using a generalized linear model with 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.

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. The regression coefficient represents the effect of an individual covariate. For example, a coefficient value of -0.5 for a comorbidity would be interpreted to mean that, on average, patients with that comorbidity had a change in self-care score that was 0.5 self-care units less than patients without that comorbidity.

Risk adjustors were added to the model together and decisions were made to retain or drop each risk adjustor based on its sample size, regression coefficient, significance level, and clinical relevance to self-care outcomes. For example, we dropped comorbidities that no longer showed a negative association with the dependent variable, because comorbidities are expected to limit functional improvement. We added comorbidities that showed a significant negative association with the dependent variable. The final risk adjustor decisions were based on a combination of clinical reasoning and statistical findings.

Risk adjustors included in the final model are described below, and also presented in S.2b. Data Dictionary, Code Table, or Value Sets.

***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 35–44 years (coefficient = 0.2755, *p* = 0.002), 45–54 years (coefficient = 0.4569, *p* < 0.001), and 55–64 years (coefficient = 0.2666, *p* < 0.001) had significantly larger change in self-care scores. Patients 75–84 years (coefficient = -0.5869,  *p* < 0.001), 85–90 years (coefficient = -1.2785, *p* < 0.001), and over 90 years (coefficient = -2.0677, *p* < 0.001) had significantly, and progressively, smaller change in self-care scores than patients in the reference category. Patients younger than 35 years did not have significantly different change scores compared with the reference category. Nevertheless, we chose not to collapse this group 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 continuous form (coefficient = 0.0793, *p* < 0.001) and the squared form (coefficient = -0.0163, *p* < 0.001) of admission self-care scores had significant effects. Thus, we included both forms of admission self-care in the final model.

***Primary Diagnosis Groups Based on IRF Primary Diagnosis:*** 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 (e.g., polyneuropathy), (8) fractures and other multiple trauma, (9) hip and knee replacements, (10) amputation, (11) other orthopedic conditions (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. The 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 (-10.1155, *p* < 0.001).

***Interaction between Primary Diagnosis Groups and Admission Self-Care Scores:*** The relationship between admission self-care and change in self-care scores may vary based on the patient’s primary diagnosis group. To account for this, 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. All interaction terms were significant, as shown in Attachment 1. All interaction terms were retained in the model.

***Prior Surgery:*** We included patients who had a major surgery during the 100 days prior to admission as a risk adjustor in the model, because patients who have recently undergone a major surgery tend to have more functional improvement than patients with medical issues without surgery (Coefficient = 0.1952, *p* < 0.001).

***Prior Functioning - Self-Care:*** We included patients’ functional ability in self-care before the 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 medical issue. 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 (coefficient = -4.3325,  *p* < 0.001) was larger than that for the “some help” (coefficient = -1.9657, *p* < 0.001) 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 or needed some help in indoor ambulation had significantly smaller change in self-care scores (coefficient = -0.8074, *p* < 0.001) 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 (full time/part time), mechanical lift, and orthotics/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 (coefficient = -1.9134, *p* < 0.001), followed by prior use of wheelchair or scooter (coefficient = -0.9244, *p* < 0.001), prior use of orthotics/prosthetics (-.4816, *p* < 0.001) and prior use of a walker (coefficient = -0.0560, *p* = 0.004).

***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 = -0.9422, *p* < 0.001) 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 = -1.3024, *p* < 0.001) compared with the reference category.

***Cognitive Function Assessed by the Brief Interview for Mental Status:*** Based on Brief Interview for Mental Status (BIMS) 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 = -0.7763, *p* < 0.001) and those with severely impaired cognitive function (coefficient = -1.7718, *p* < 0.001) 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 = -0.9358, *p* < 0.001) 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 (e.g., renal failure).” Patients with bladder incontinence (coefficient = -1.2459, *p* < 0.001) 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 = -1.1527, *p* < 0.001).

***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 = -2.3437, *p* < 0.001) having a larger negative coefficient compared with the “less than daily” or “daily incontinence” category (coefficient = -0.8470, *p* < 0.001).

***Swallowing Ability:*** Our model included two separate risk adjustors related to patients’ swallowing ability: (1) need for modified food consistency, 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 = -0.7813, p < 0.001) having a larger coefficient than the modified food consistency group (coefficient = -0.6785, p < 0.001).

***Low Body Mass Index (BMI):*** We included a risk adjustor for patients with low BMI based on their height and weight. Patients with low BMI had significantly smaller change in self-care scores (coefficient= -0.3363, p < 0.001) compared with the reference category.

***Comorbidities:*** We used all the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes reported on the IRF-PAI (Item 24 - Comorbid Conditions) to identify patient comorbidities. ICD-10-CM codes were used to assign patients into one or more of the Hierarchical Condition Categories. We tested approximately 135 of the Hierarchical Condition Categories that were determined to be clinically relevant to self-care outcomes.

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.”

The attached model in S.2b. Data Dictionary, Code Table, or Value Sets 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 certain cancers; dementia; tetraplegia; Amyotrophic Lateral Sclerosis and other motor neuron diseases; Cerebral Palsy; legally blind; and major fracture, except of skull, vertebrae, or hip.

**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)**

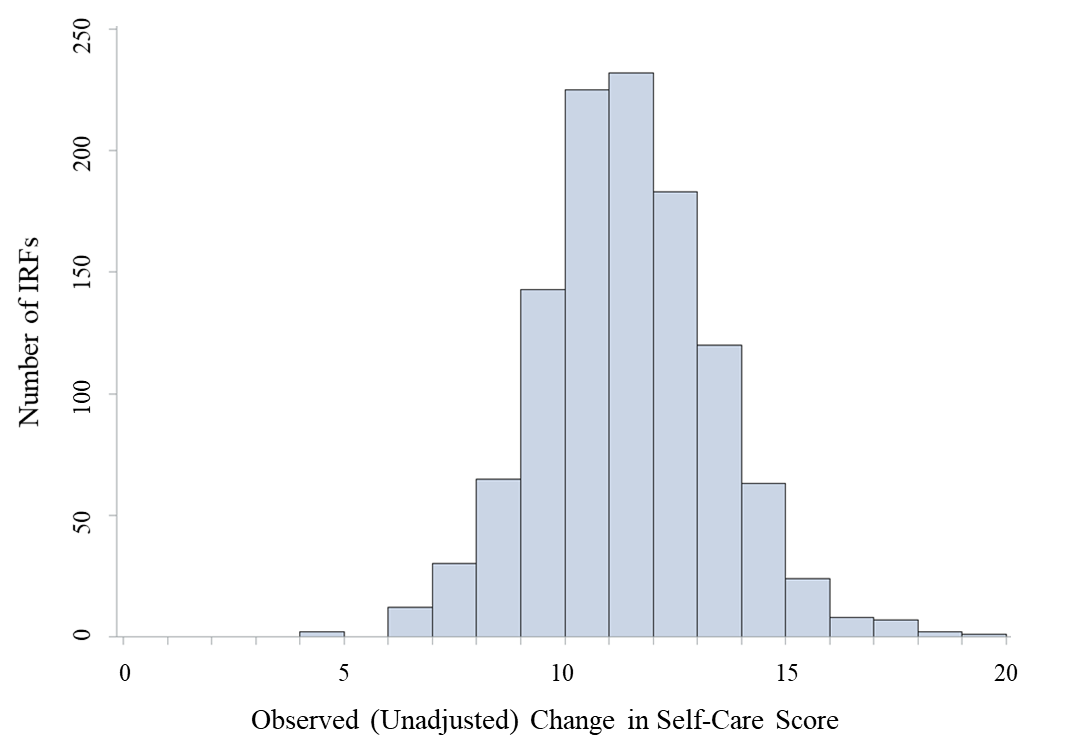
**2b3.4a. What were the statistical results of the analyses used to select risk factors?**

Results of the final risk adjustment model are shown in S.2b. Data Dictionary, Code Table, or Value Sets, along with regression coefficients and significance values of the final set of risk adjustors.

As described above, decisions were made to retain or drop each risk adjustor based on its sample size, regression coefficient, significance level, and clinical relevance to self-care outcomes. For example, we dropped comorbidities that no longer showed a negative association with the dependent variable, because comorbidities are expected to limit functional improvement. We added comorbidities that showed a significant negative association with the dependent variable. The final risk adjustor decisions were based on a combination of clinical reasoning and statistical findings.

The overall model was a significant predictor of change in self-care scores, with a *p*-value less than 0.001. The overall model R-square was 0.23, indicating that 23% 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.

Distributions of the facility-level mean unadjusted and risk adjusted change in self-care scores are shown in **Figures 2 and 3 and Table 12.** 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 (n=1,117)**

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV50)

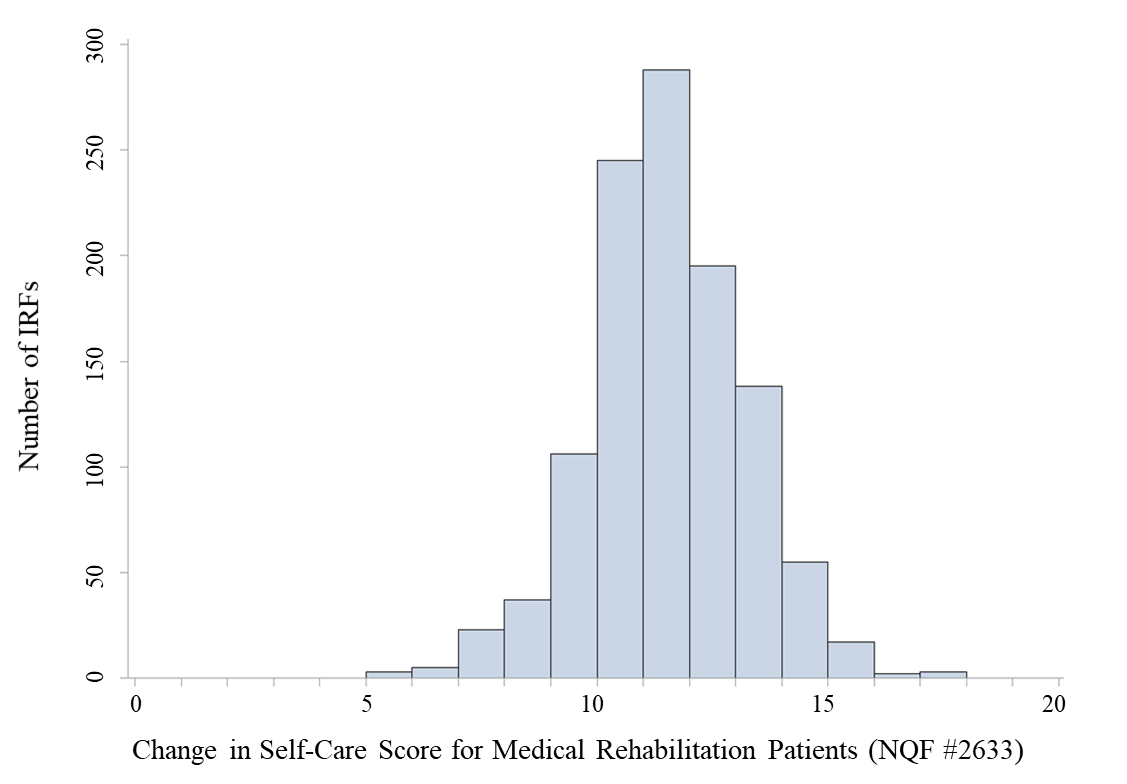
**Figure 2 Data Table. Distribution of Facility-Level Mean Unadjusted Change in Self-Care Scores (n=1,117)**

| **Observed (Unadjusted) Change in Self-Care Scores\*** | **Number of IRFs** |
| --- | --- |
| 4.0 | 2 |
| 6.0 | 6 |
| 7.0 | 20 |
| 8.0 | 45 |
| 9.0 | 93 |
| 10.0 | 178 |
| 11.0 | 251 |
| 12.0 | 231 |
| 13.0 | 132 |
| 14.0 | 91 |
| 15.0 | 45 |
| 16.0 | 9 |
| 17.0 | 9 |
| 18.0 | 4 |
| 20.0 | 1 |
| **Total** | **1117** |

\*Scores were rounded to the nearest whole number for the figure

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV50)

**Figure 3. Distribution of Facility-Level Mean Risk Adjusted Change in Self-Care Scores (n=1,117)**

****

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV50)

**Figure 3 Data Table. Distribution of Facility-Level Mean Risk Adjusted Change in Self-Care Scores (n=1,117)**

| **Facility-Level Mean Risk Adjusted Change in Self-Care Scores\*** | **Number of IRFs** |
| --- | --- |
| 5.0 | 2 |
| 6.0 | 3 |
| 7.0 | 11 |
| 8.0 | 34 |
| 9.0 | 63 |
| 10.0 | 179 |
| 11.0 | 266 |
| 12.0 | 262 |
| 13.0 | 159 |
| 14.0 | 98 |
| 15.0 | 27 |
| 16.0 | 8 |
| 17.0 | 4 |
| 18.0 | 1 |
| **Total** | **1,117** |

\*Scores were rounded to the nearest whole number for the figure

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV50)

Skewness and kurtosis values of the facility-level mean unadjusted change scores are larger than those of the mean risk adjusted change scores (**Table 12**), indicating that the unadjusted scores deviate from a normal distribution to a larger extent than the risk adjusted scores (i.e., performance measure scores). **Table 12** shows that the standard deviation of the mean risk adjusted change scores is slightly smaller than those of the unadjusted change scores. The mean risk adjusted change scores have a range of 5.4 to 17.5, and an interquartile range of 2.20. In contrast, the mean unadjusted change scores have a wider range of 4.2 to 19.6, and an interquartile range of 2.41.

**Table 12. Distribution of Facility-Level Mean Unadjusted and Risk Adjusted Change in Self-Care Scores (n=1,117)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Change in Self-Care Score** | **N** | **Mean (SD)** | **SE** | **Min** | **10th Pctl** | **25th Pctl** | **Median** | **75th Pctl** | **90th Pctl** | **Max** | **Skewness** | **Kurtosis** |
| Unadjusted (Observed) | 1,117 | 11.4 (2.0) | 0.1 | 4.2 | 9.1 | 10.1 | 11.3 | 12.6 | 14.0 | 19.6 | 0.2 | 0.7 |
| Risk Adjusted Performance Measure | 1,117 | 11.5 (1.7) | 0.1 | 5.4 | 9.4 | 10.4 | 11.5 | 12.6 | 13.7 | 17.5 | -0.1 | 0.6 |

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

Note: Providers with <20 stays during the 12-month testing period are excluded from facility-level analyses.

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV50)

**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.**

We examined whether 5 social risk factors affected computed performance measure scores: 1) dual eligibility (patient-level variable); 2) race/ethnicity (patient-level variable); 3) living alone (patient-level variable); 4) urbanicity based on the patient’s residence (community-level variable); and 5) socioeconomic status (SES) (community-level variable).

We obtained patients’ dual eligibility status from the Integrated Data Repository (IDR), and race/ethnicity and living alone status from the IRF-PAI. Urbanicity was determined by cross-walking beneficiary residence ZIP codes (from the IRF-PAI) to Federal Information Processing Standard Publication (FIPS) codes,[[4]](#footnote-4) then cross-walking FIPS codes to Rural-Urban Commuting Area Codes (RUCA\_2013).[[5]](#footnote-5) Socioeconomic status was determined using the Agency of Healthcare Research and Quality’s SES Index[[6]](#footnote-6) calculated based on beneficiary residence ZIP Code Tabulation Area (ZCTA). ZCTA was found by cross-walking the beneficiary residence ZIP code with ZCTA. We used data from the 2016 American Community Survey (5-year file) to calculate AHRQ SES Index, with higher values indicating higher SES.

We conducted the following analyses to examine the effect of the 5 social risk factors:

* We calculated the percentage of stays for each social risk factor subgroup;
* We calculated the change in self-care score for each social risk factor subgroup;
* We added indicators for each social risk factor to our risk adjustment model and estimated the coefficients of these risk adjusters in the model; and
* We calculated the difference in provider scores with and without social risk factor adjustment.

**Table 13** shows the distribution of the social risk factors in the calendar 2017 IRF data and the mean change in self-care score by social risk factor subgroup. We found that 12.2% of patients were dual eligible with full Medicaid benefits, 79.4% of patients were white, and 29.7% were living alone. We also found that 83.8% of IRF patients lived in urban areas. The lowest quartile of AHRQ SES index ranged from 27.9 - 49.5; the highest quartile ranged from 55.3 – 75.7.

The mean unadjusted change in self-care score varied by dual eligibility status, race, and living alone status. Patients who were dual eligible with full Medicaid benefits had on average 11.0 units of change in self-care while patients who were dual eligible without full Medicaid benefits or who were non-dual eligible had on average 12.0 and 11.6 units of change in self-care, respectively. For race, the highest mean change in self-care during 2017 was found among patients who were white (11.6 units of change), multiracial (11.5 units of change), or Native American or Alaskan Native (11.4 units of change) whereas the lowest was among patients who were Asian (10.4 units of change). Patients who were living alone prior to their hospitalization had on average 12.0 units of change in self-care whereas those not living alone had 11.3 units of change in self-care. The mean unadjusted change in self-care scores were similar across Hispanic ethnicity, urbanicity, and SES. Patients who were Hispanic had a similar mean change in self-care score (11.8 units of change) as patients who were of non-Hispanic ethnicity (11.5 units of change). The average unit of change in self-care for patients who were living in rural and urban locations, ranged from 11.5 to 11.7 units of change in self-care, and by AHRQ SES Index, ranged from to 11.2 units of change for quartile 4 to 11.7 units of change for quartile 1.

**Table 13. Distribution of Social Risk Factors and Mean Change in Self-Care Score for IRF Patients (N = 428,710)**

| **Social Risk Factor** | **n (%)** | **Observed Change in Self-Care (unadjusted)** |
| --- | --- | --- |
| ***Dual Eligibility*** |  |  |
| Dual with full Medicaid | 52,450 (12.2) | 11.0 |
| Dual without full Medicaid | 25,113 (5.9) | 12.0 |
| Non-dual | 351,147 (81.9) | 11.6 |
| ***Race*** |  |  |
| White | 340,398 (79.4) | 11.6 |
| Black | 46,949 (11.0) | 10.9 |
| Asian | 6,689 (1.6) | 10.4 |
| American Indian or Alaska Native | 1,339 (0.3) | 11.4 |
| Native Hawaiian or Pacific Islander | 1,546 (0.4) | 10.7 |
| Multiracial | 246 (0.1) | 11.5 |
| Missing | 31,543 (7.4) | 11.8 |
| ***Hispanic Ethnicity*** |  |  |
| Yes | 20,147 (4.7) | 11.8 |
| No | 408,563 (95.3) | 11.5 |
| ***Living Alone*** |  |  |
| Yes | 127,218 (29.7) | 12.0 |
| No | 301,492 (70.3) | 11.3 |
| ***Urbanicity*** |  |  |
| Urban | 359,388 (83.8) | 11.5 |
| Suburban | 48,965 (11.4) | 11.8 |
| Rural | 18,000 (4.2) | 11.7 |
| Missing | 2,357 (0.5) | 10.9 |
| ***AHRQ SES Index\**** |  |  |
| Quartile 1 (27.9 - 49.5) | 106,256 (24.8) | 11.7 |
| Quartile 2 (49.5 – 52.1) | 106,438 (24.8) | 11.7 |
| Quartile 3 (52.1 – 55.3) | 106,876 (24.9) | 11.5 |
| Quartile 4 (55.3 – 75.7) | 107,203 (25.0) | 11.2 |
| Missing | 1,937 (0.5) | 10.8 |

\* based on beneficiary residence. AHRQ = Agency for Healthcare Research.

Notes: N= number of patient stays; patient-level exclusion criteria applied; unadjusted Change in Self-Care values are reported as units of change in self-care (possible range: -35 to 35).

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: LP65)

**Table 14** shows the social risk factor estimates in our Generalized Linear regression model. Dual eligibility patients with full Medicaid benefits had lower self-care changes scores (coefficient = -0.3760, *p* < 0.001) while patients with partial Medicaid benefits had higher self-care change scores (coefficient =0.2438, *p* < 0.001), on average, than patients who were non-dual eligible. Compared to patients who were White, Black patients (coefficient = -0.5551, *p* < 0.001), Asian patients (coefficient = -0.4772, *p* < 0.001), and patients of Native Hawaiian or Pacific Islander descent (coefficient = -0.4248, *p* = 0.003) had lower self-care changes, on average. Hispanic patients (coefficient = 0.1771, *p* = 0.004) had higher self-care change scores than non-Hispanic patients. Patients who lived alone (coefficient = 0.4732, *p* < 0.001) had higher self-care change scores than patients who did not live alone prior to their hospitalization. Patients living in rural areas had similar change in self-care scores compared with patients living in urban areas. Patients residing in AHRQ SES Index quartiles 1-3 had higher self-care change scores, on average, than patients residing in AHRQ SES Index quartile 4.

**Table 14. Effect of Social Risk Factors in the IRF Change in Self-Care Regression Model (N = 428,192)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Social Risk Factor** | **Estimate** | **SE** | ***p*-value** |
| ***Dual Eligibility (reference =*** ***Non-dual)*** | | | |
| Dual with full Medicaid | -0.3760 | 0.03 | <.001 |
| Dual without full Medicaid | 0.2438 | 0.04 | <.001 |
| ***Race/Ethnicity (reference = White)*** | | | |
| Black | -0.5551 | 0.03 | <.001 |
| Asian | -0.4772 | 0.07 | <.001 |
| American Indian or Alaska Native | -0.1013 | 0.16 | 0.514 |
| Native Hawaiian or Pacific Islander | -0.4248 | 0.14 | 0.003 |
| Multiracial | 0.1038 | 0.36 | 0.774 |
| Missing | -0.0252 | 0.05 | 0.610 |
| ***Hispanic Ethnicity*** | 0.1771 | 0.06 | 0.004 |
| ***Living Alone*** | 0.4732 | 0.02 | <.001 |
| ***Urbanicity\* (reference = Urban)*** | | | |
| Rural | -0.0466 | 0.04 | 0.292 |
| Suburban | 0.0296 | 0.03 | 0.298 |
| Missing | -0.2512 | 0.28 | 0.361 |
| ***AHRQ SES Index\* (reference = Quartile 4 (55.6 to 75.7))*** | | | |
| Quartile 1 (28.9 to 49.6) | 0.5516 | 0.03 | <.001 |
| Quartile 2 (49.7 to 52.2) | 0.3961 | 0.03 | <.001 |
| Quartile 3 (52.3 to 55.5) | 0.2917 | 0.02 | <.001 |
| Missing | 0.0140 | 0.75 | 0.985 |

\* based on patient residence. AHRQ = Agency for Healthcare Research.

Note: SE=Standard error; Patient-level exclusion criteria applied.

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: LP65)

**Table 15** shows the distribution of the change in self-care performance measure scores with and without social risk factor adjustment. Overall, social risk factor adjustment had minimal impact on providers’ calculated performance measure scores. Between the two sets of scores, the difference in mean scores was 0.0 units of change in self-care, with a standard deviation of 0.2 and interquartile range of 0.2.

**Table 15: Distribution of IRF Change in Self-Care Scores With and Without Adjustment for Social Risk Factors (n = 1,117)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Change in Self-Care Scores** | **Mean** | **SD** | **Min** | **25th Pct** | **Median** | **75th Pct** | **Max** |
| Not adjusting for SRF | 11.5 | 1.7 | 5.4 | 10.4 | 11.5 | 12.6 | 17.5 |
| Adjusting for SRF | 11.5 | 1.7 | 5.3 | 10.4 | 11.5 | 12.5 | 17.7 |
| Difference in Scores (SRF-adjusted minus non-SRF adjusted scores)\* | 0.0 | 0.2 | -0.4 | -0.1 | 0.0 | 0.1 | 0.6 |

\*Calculated as SRF-adjusted score minus non-SRF adjusted score for each facility.

Note: SD=Standard deviation; Min=minimum score; Max=maximum score; Pct = percentile. SRF = social risk factors.

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: LP65)

Our testing of social risk factors and their relationship to patients’ change in self-care scores indicate that some factors (full dual eligibility, Black, Asian or Native Hawaiian race) were tied to lower self-care change scores while others (lower SES, living alone, Hispanic ethnicity) were tied to higher self-care change scores. Although race and dual eligibility were associated with lower changes in self-care scores the effects were small, and we believe that further study is needed to better understand how social risk factors can influence health outcomes. Our risk adjustment model explained 23% of variance in change in self-care, and the inclusion of these five social risk factors did not explain any additional variance (r-squared = 0.234). In addition, the mean and median Change in Self-Care Score with and without adjusting for the social risk factors are the same.

As noted in the Assistant Secretary for Planning and Evaluation’s Report to Congress entitled “*Social Risk Factors Performance under Value-Based Purchasing*” (<https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs>), adjusting performance measures for social factors may mask disparities in the quality of care provided, which could reduce the ability to identify and reduce them. In addition, when differences in quality are related to poor performance, bias, or discrimination, adjusting performance measures could excuse the delivery of worse care to beneficiaries with social risk factors.

Therefore, we do not adjust for social risk factors in our risk adjustment model for the IRF Change in Self-Care performance measure. We will continue to monitor the impact of social risk factors on providers’ performance measure scores.

**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*)  
*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#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, and outlier influence.

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

Overall, the model explained 23% of variance in change in self-care.

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

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 0.0013.

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

To assess model performance and stability across the sample, we divided our dataset into deciles of expected values and calculated the ratio of average expected change score to average observed change score within each decile. A ratio of 1 would indicate perfect agreement between average expected and observed change scores. We expect that the risk adjusted model performance will be stable among IRFs regardless of whether they have patients with low or high change scores on average.

As seen in **Table 16**, the average expected to observed change score ratios within each decile approximated 1.2, with a range of 0.9 to 1.3, validating model performance. There was little variability in average expected to observed change score ratios across deciles, supporting model stability across the range of expected change scores and across the sample.

**Table 16. Ratio of Average Expected to Average Observed Change in Self-Care Scores Across Deciles of Expected Change Scores**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Deciles of Expected Change Scores** | **N** | **Average Expected Change Score** | **Average Observed Change Score** | **Average Expected to Observed Ratio** |
| Decile 1 (-6.4 – 7.5) | 42,819 | 5.7 | 6.2 | 0.9 |
| Decile 2 (7.5 – 9.0) | 42,819 | 8.3 | 8.3 | 1.1 |
| Decile 3 (9.0 – 10.0) | 42,819 | 9.5 | 9.3 | 1.2 |
| Decile 4 (10.0 – 10.9) | 42,820 | 10.5 | 10.3 | 1.3 |
| Decile 5 (10.9 – 11.8) | 42,819 | 11.3 | 11.1 | 1.3 |
| Decile 6 (11.8 – 12.5) | 42,819 | 12.2 | 12.0 | 1.3 |
| Decile 7 (12.5 – 13.3) | 42,820 | 12.9 | 12.8 | 1.3 |
| Decile 8 (13.3 – 14.2) | 42,819 | 13.8 | 13.7 | 1.3 |
| Decile 9 (14.2 – 15.4) | 42,819 | 14.8 | 14.8 | 1.2 |
| Decile 10 (15.4 – 23.6) | 42,819 | 16.5 | 17.1 | 1.2 |
| **Total Sample** | 428,192 | 11.5 | 11.5 | 1.2 |

Note: N = number of patient stays; Observed Change in Self-Care values are reported as units of change in self-care (possible range: -35 to 35); Providers with < 20 stays during the 12-month testing period are excluded.

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV50)

**2b3.9. Results of Risk Stratification Analysis**:

Not applicable – no stratification

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

**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*)

None

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

**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)*

For the IRF Change in Self-Care Score performance measure, we examined whether each IRF’s calculated performance measure score (i.e., the risk-adjusted change in self-care score) was worse than, better than, or the same as national average performance of all IRFs. For each IRF, we calculated the 95% confidence interval for the computed performance measure score and compared this with the national mean observed change score. Facilities whose confidence interval was lower than the national mean observed change score were considered to have worse performance than the national average. Facilities whose confidence interval was higher than the national mean observed change score were considered to have better performance than the national average. Facilities whose confidence interval overlapped with the national mean observed change score were considered to be similar to national average performance.

**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*)

**Table 17** shows that for the IRF Change in Self-Care Score measure, 33.0% of IRFs had 95% confidence intervals lower than the national mean change score, indicating worse than national average performance. As shown in **Figure 3** above, the IRF calculated performance scores (i.e., the risk-adjusted change in self-care scores) are generally normally distributed.

**Table 17. Comparison of Facility-Level Measure Scores with National Average Performance for Change in Self-Care Score (N = 1,117)**

|  |  |  |  |
| --- | --- | --- | --- |
| Measure Name | Facility Performance Worse than National Average  N (%) | Facility Performance Better than National Average  N (%) | Facility Performance Same as National Average  N (%) |
| Change in Self-Care Score | 369 (33.0%) | 341 (30.5%) | 407 (36.4%) |

Note: Providers with < 20 stays during the 12-month testing period are excluded.

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV53)

**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?*)

These results demonstrate the ability of the measures to discriminate among facilities based on facility-level measure performance.

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

**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*)

Not applicable

**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*)

Not applicable

**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*)

Not applicable

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

**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*)

We ran frequencies of missing data for each self-care data element at admission and discharge as well as each of the risk adjustors after applying the exclusion criteria to examine the extent and distribution of missing data. Missing data on the IRF-PAI is identified as a dash (-), which is coded by providers to indicate they have “No information.” Dash use is expected to be a rare occurrence and coding guidance is provided through in-person and web-based trainings, training manuals, and responses to help desk inquiries.

**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*)

The frequencies of missing data for each self-care data element using data from the IRF-PAI are reported in **Table 18** at admission and discharge. Across all self-care data elements, at admission and discharge, the number of cases in which the data element data are missing is very low – less than 0.1%.

**Table 18. Self-Care Data Elements: Missing Data (n=427,639)**

|  |  |  |
| --- | --- | --- |
|  | **Admission:  Not Assessed (-)** | **Discharge:**  **Not Assessed (-)** |
| **Self-Care Data Elements** |  |  |
| GG0130A: Self-Care - Eating | 31 (< 0.01%) | 58 (< 0.01%) |
| GG0130B: Self-Care - Oral Hygiene | 56 (<0.01%) | 56 (< 0.01%) |
| GG0130C: Self-Care - Toileting Hygiene | 65 (<0.01%) | 39 (< 0.01%) |
| GG0130E: Self-Care - Shower/Bathe Self | 88 (< 0.01%) | 62 (< 0.01%) |
| GG0130F: Self-Care - Upper Body Dressing | 39 (< 0.01%) | 42 (< 0.01%) |
| GG0130G: Self-Care - Lower Body Dressing | 28 (< 0.01%) | 36 (< 0.01%) |
| GG0130H: Self-Care - Putting on/Taking Off Footwear | 47 (< 0.01%) | 70 (< 0.01%) |
| Total | 354 (< 0.1%) | 363 (< 0.1%) |

Source: RTI analysis of IRF-PAI, January – December 2017. (Program reference: MV45).

The frequencies of missing data for each of the risk adjustors (available upon request) is also very low, ranging from no missing data for Age and Primary Diagnosis to 0.1% for the BIMS. Though missing data is rare, it is still accounted for in the calculation of the risk adjustors. For example, when determining Prior Surgery from the J2000 data element, a dash (-) on the IRF-PAI is recoded to “0” to indicate no Prior Surgery rather than dropping the patient from the performance measure calculation.

**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*)

There is a very small percentage of cases with missing data, and we believe this very small percentage is unlikely to cause significant bias.

**Appendix A**

**Table A-1. Intercept and Risk-Adjustor Definitions and Covariate Values for the Change in Self-Care Measure, (NQF #2633)**

| **Risk Adjustor** | **Risk Adjustor Category** | **IRF-PAI Item(s) and Calculations** | **Intercept and Coefficients for Change in Self-Care (NQF #2633)**  **All values have 4 decimal places** |
| --- | --- | --- | --- |
| Intercept | -- | -- | 23.6369 |
| Age Group | <35 years | Truncate (Item 12 – Item 6) = age; If age <35 years = 1; else = 0 | 0.0363 |
| Age Group | 35-44 years | Truncate (Item 12 – Item 6) = age; If age 35–44 years = 1; else = 0 | 0.2755 |
| Age Group | 45-54 years | Truncate (Item 12 – Item 6) = age; If age 45–54 years = 1; else = 0 | 0.4569 |
| Age group | 55-64 years | Truncate (Item 12 – Item 6) = age; If age 55–64 years = 1; else = 0 | 0.2666 |
| Age Group | 75-84 years | Truncate (Item 12 – Item 6) = age; If age 75–84 years = 1; else = 0 | -0.5869 |
| Age Group | 85-90 years | Truncate (Item 12 – Item 6) = age; If age 85–90 years = 1; else = 0 | -1.2785 |
| Age Group | >90 years | Truncate (Item 12 – Item 6) = age; If age >90 years = 1; else = 0 | -2.0677 |
| Admission Self-Care - continuous form | Admission Self-Care - continuous form | Admission Self-Care Score = (GG0130A1 + GG0130B1 + GG0130C1 + GG0130E1 + GG0130F1 + GG0130G1 + GG0130H1) | 0.0793 |
| Admission Self-Care - squared form | Admission Self-Care - squared form | Admission Self-Care Score Squared = (GG0130A1 + GG0130B1 + GG0130C1 + GG0130E1 + GG0130F1 + GG0130G1 + GG0130H1) \* (GG0130A1 + GG0130B1 + GG0130C1 + GG0130E1 + GG0130F1 + GG0130G1 + GG0130H1) | -0.0163 |
| Primary Diagnosis Group | Stroke | = 1 if Item 21A = 0001.1 or 0001.2 or 0001.3 or 0001.4 or 0001.9; else = 0 | -9.6701 |
| Primary Diagnosis Group | Non-Traumatic Brain Dysfunction | = 1 if Item 21A = 0002.1 or 0002.9; else = 0 | -5.7268 |
| Primary Diagnosis Group | Traumatic Brain Dysfunction | = 1 if Item 21A = 0002.21 or 0002.22 or 0014.1 or 0014.2; else = 0 | -4.1526 |
| Primary Diagnosis Group | Non-Traumatic Spinal Cord Dysfunction | = 1 if Item 21A = 0004.110 or 0004.111 or 0004.112 or 0004.120 or 004.1211 or 0004.1212 or 0004.130; else = 0 | -6.2803 |
| Primary Diagnosis Group | Traumatic Spinal Cord Dysfunction | = 1 if Item 21A = 0004.210 or 0004.211 or 0004.212 or 0004.220 or 004.2211 or 0004.2212 or 0004.230 or 0014.3; else = 0 | -10.1155 |
| Primary Diagnosis Group | Progressive Neurological Conditions | = 1 if Item 21A = 0003.1 or 0003.2; else = 0 | -7.0579 |
| Primary Diagnosis Group | Other Neurological Conditions | = 1 if Item 21A = 0003.3 or 0003.4 or 0003.5 or 0003.8 or 0003.9; else = 0 | -4.5214 |
| Primary Diagnosis Group | Fractures and Other Multiple Trauma | = 1 if Item 21A = 0008.11 or 0008.12 or 0008.2 or 0008.3 or 0008.4 or 0014.9; else = 0 | -5.5931 |
| Primary Diagnosis Group | Amputation | = 1 if Item 21A = 0005.1 or 0005.2 or 0005.3 or 0005.4 or 0005.5 or 0005.6 or 0005.7 or 0005.9; else = 0 | -6.9618 |
| Primary Diagnosis Group | Other Orthopedic Conditions | = 1 if Item 21A = 0006.1 or 0006.2 or 0006.9 or 0007.1 or 0007.2 or 0007.3 or 0007.9 or 0008.9; else = 0 | -5.7078 |
| Primary Diagnosis Group | Debility, Cardiorespiratory Conditions | = 1 if Item 21A = 0009 or 0010.1 or 0010.9 or 0016 or 0017.4 or 0017.51 or 0017.52; else = 0 | -4.2683 |
| Primary Diagnosis Group | Medically Complex Conditions | = 1 if Item 21A = 0011 or 0012.1 or 0012.9 or 0013 or 0015 or 0017.1 or 0017.2 or 0017.31 or 0017.32 or 0017.6 or 0017.7 or 0017.8 or 0017.9; else = 0 | -4.7541 |
| Interaction of admission self-care score and primary diagnosis group | Stroke | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Stroke (see above) | 0.2992 |
| Interaction of admission self-care score and primary diagnosis group | Non-Traumatic Brain Dysfunction | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Non-Traumatic Brain Dysfunction (see above) | 0.1598 |
| Interaction of admission self-care score and primary diagnosis group | Traumatic Brain Dysfunction | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Traumatic Brain Dysfunction (see above) | 0.0986 |
| Interaction of admission self-care score and primary diagnosis group | Non-Traumatic Spinal Cord Dysfunction | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Non-Traumatic Spinal Cord Dysfunction (see above) | 0.2025 |
| Interaction of admission self-care score and primary diagnosis group | Traumatic Spinal Cord Dysfunction | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Traumatic Spinal Cord Dysfunction (see above) | 0.3498 |
| Interaction of admission self-care score and primary diagnosis group | Progressive Neurological Conditions | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Progressive Neurological Conditions (see above) | 0.2097 |
| Interaction of admission self-care score and primary diagnosis group | Other Neurological Conditions | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Other Neurological Conditions (see above) | 0.1601 |
| Interaction of admission self-care score and primary diagnosis group | Fractures and Other Multiple Trauma | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Fractures and Other Multiple Trauma (see above) | 0.1739 |
| Interaction of admission self-care score and primary diagnosis group | Amputation | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Amputation (see above) | 0.2112 |
| Interaction of admission self-care score and primary diagnosis group | Other Orthopedic Conditions | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Other Orthopedic Conditions (see above) | 0.1773 |
| Interaction of admission self-care score and primary diagnosis group | Debility, Cardiorespiratory Conditions | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Debility, Cardiorespiratory Conditions (see above) | 0.1448 |
| Interaction of admission self-care score and primary diagnosis group | Medically Complex Conditions | Admission self-care: continuous form (see above) multiplied by Primary diagnosis: Medically Complex Conditions (see above) | 0.1533 |
| Prior surgery | Surgical | =1 if J2000 = 1; else = 0 | 0.1952 |
| Prior functioning: self-care | Dependent | =1 if GG0100A = 1; else = 0 | -4.3325 |
| Prior functioning: self-care | Some help | =1 if GG0100A = 2; else = 0 | -1.9657 |
| Prior functioning: indoor ambulation (combined) | Dependent, Some help | =1 if GG0100B = 1 or GG0100B = 2; else = 0 | -0.8074 |
| Prior Mobility Device/Aid | Walker | =1 if GG0110D = 1; else = 0 | -0.0560 |
| Prior Mobility Device/Aid | Wheelchair/Scooter Full Time/Part Time | =1 if GG0110A = 1 or GG0110B = 1; else = 0 | -0.9244 |
| Prior Mobility Device/Aid | Mechanical Lift | =1 if GG0110C =1; else = 0 | -1.9134 |
| Prior Mobility Device/Aid | Orthotics/Prosthetics | =1 if GG0110E = 1; else = 0 | -0.4816 |
| Stage 2 Pressure Ulcer | Present | =1 if M0300B1 ≥ 1; else = 0 | -0.9422 |
| Stage 3, 4 or Unstageable Pressure Ulcer | Present | =1 if M0300C1 ≥ 1 or M0300D1 ≥ 1 or M0300E1 ≥ 1 or M0300F1 ≥ 1 or M0300G1 ≥ 1; else = 0 | -1.3024 |
| Cognitive Function: Brief Interview for Mental Status score | Moderately Impaired | =1 if C0500 = 8, 9, 10, 11, or 12 or ([C0900A = 1 and C0900B = 1] or [C0900B = 1 and C0900C = 1] or [C0900A = 1 and C0900C = 1]) or [C0900A = 1 and C0900E = 1] or [C0900B = 1 and C0900E = 1] or [C0900C = 1 and C0900E = 1]); else = 0 | -0.7763 |
| Cognitive Function: Brief Interview for Mental Status score | Severely Impaired | =1 if C0500 = ≤ 7 or (C0900Z = 1 or ([C0900A=1 and C0900B = 0, and C0900C = 0, and C0900E = 0] or [C0900B=1 and C0900A = 0, and C0900C = 0, and C0900E = 0] or [C0900C=1 and C0900A = 0, and C0900B = 0, and C0900E = 0] or [C0900E=1 and C0900A = 0, and C0900B = 0, and C0900C = 0]); else = 0 | -1.7718 |
| Communication Impairment | Moderate to Severe | =1 if BB0800 = 1 or BB0800 = 2 or BB0700 = 1 or BB0700 = 2; else = 0 | -0.9358 |
| Bladder Incontinence | Less than daily, Daily, Always incontinent | =1 if H0350 = 2 or H0350 = 3 or H0350 = 4; else = 0 | -1.2459 |
| Bladder Incontinence | Urinary catheter | =1 if H0350 = 9; else = 0 | -1.1527 |
| Bowel Incontinence | Always incontinent | =1 if H0400 = 3; else = 0 | -2.3437 |
| Bowel Incontinence | Less than daily, Daily | =1 if H0400 = 1 or H0400 = 2; else = 0 | -0.8470 |
| Swallowing Ability | Modified Food Consistency | =1 if K0110B = 1; else = 0 | -0.6785 |
| Swallowing Ability | Tube/Parenteral Feeding | =1 if K0110C = 1; else = 0 | -0.7813 |
| Body Mass Index (BMI) | Low BMI | = 1 if BMI ≥ [12.0] AND ≤ [19.0]; = 0 if BMI < [12.0] OR BMI > [19.0]; = 0 if Item 25A = [0, 00, -] OR Item 26A = [-]; else = 0. Where: BMI = (([Item 26A] \* 703) / Item 25A2) and the resulting value is rounded to one decimal place. | -0.3363 |
| Comorbidity Condition Group 1 | Viral and Late Effects Central Nervous System Infections (HCC4) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC4; =0 if Item 21A = 0017.1 or 0002.1 or 0002.9 or 0004.11 thru 0004.13; else = 0 | -0.2481 |
| Comorbidity Condition Group 2 | Other Infectious Diseases (HCC7) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #7; =0 if Item 21A = 0017.1; else = 0 | -0.3900 |
| Comorbidity Condition Group 3 | Metastatic Cancer and Acute Leukemia (HCC8) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #8; =0 if Item 21A = 0017.2; else = 0 | -1.1700 |
| Comorbidity Condition Group 4 | Lung and Other Severe Cancers (HCC9) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #9; =0 if Item 21A = 0017.2; else = 0 | -0.6272 |
| Comorbidity Condition Group 5 | Lymphoma and Other Cancers (HCC10) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #10; =0 if Item 21A = 0017.2; else = 0 | -0.4787 |
| Comorbidity Condition Group 6 | Diabetes: Diabetes with Chronic Complications (HCC18) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #18; =0 if Item 21A = 0017.31, 0017.32; else = 0 | -0.0736 |
| Comorbidity Condition Group 7 | Diabetes without Complication (HCC19) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #19; =0 if Item 21A = 0017.31, 0017.32; else = 0 | -0.1306 |
| Comorbidity Condition Group 8 | Protein-Calorie Malnutrition (HCC21) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #21; =0 if Item 21A = 0017.31, 0017.32; else = 0 | -0.2603 |
| Comorbidity Condition Group 9 | Other Significant Endocrine and Metabolic Disorders (HCC23) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #23; else = 0 | -0.1020 |
| Comorbidity Condition Group 10 | Intestinal Obstruction/Perforation (HCC33) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #33; else = 0 | -0.2264 |
| Comorbidity Condition Group 11 | Bone/Joint/Muscle Infections/Necrosis (HCC39) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #39; =0 if Item 21A = 0017.1, 0017.7; else = 0 | -0.5718 |
| Comorbidity Condition Group 12 | Delirium and Encephalopathy (HCC50) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #50; else = 0 | -0.4520 |
| Comorbidity Condition Group 13 | Dementia: Dementia With Complications (HCC51) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #51; =0 if Item 21A = 0002.1, 0002.9; else = 0 | -1.3119 |
| Comorbidity Condition Group 14 | Dementia Without Complications (HCC52) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #52; =0 if Item 21A = 0002.1, 0002.9; else = 0 | -1.1398 |
| Comorbidity Condition Group 15 | Nonpsychotic Organic Brain Syndromes/Conditions (HCC53) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #53; else = 0 | -0.2646 |
| Comorbidity Condition Group 16 | Reactive and Unspecified Psychosis (HCC59) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #59; else = 0 | -0.3631 |
| Comorbidity Condition Group 17 | Tetraplegia (HCC70)\* | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #70; =0 if Primary Diagnosis Group = Non-traumatic spinal cord dysfunction or Traumatic spinal cord dysfunction; else = 0 | -2.1428 |
| Comorbidity Condition Group 18 | Paraplegia (HCC71) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #71; =0 if Primary Diagnosis Group = Non-traumatic spinal cord dysfunction or Traumatic spinal cord dysfunction; else = 0 | -0.8584 |
| Comorbidity Condition Group 19 | Spinal Cord Disorders/Injuries (HCC72) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #72; =0 if Primary Diagnosis Group = Non-traumatic spinal cord dysfunction or Traumatic spinal cord dysfunction; else = 0 | -0.4498 |
| Comorbidity Condition Group 20 | Amyotrophic Lateral Sclerosis and Other Motor Neuron Disease (HCC73) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #73; =0 if Item 21A = 0003.8, 0003.9; else = 0 | -1.6164 |
| Comorbidity Condition Group 21 | Cerebral Palsy (HCC74) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #74; =0 if Item 21A = 0003.5; else = 0 | -2.1701 |
| Comorbidity Condition Group 22 | Multiple Sclerosis (HCC77) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #77; =0 if Item 21A = 0003.1; else = 0 | -0.6123 |
| Comorbidity Condition Group 23 | Parkinson's and Huntington's Diseases (HCC78) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #78; =0 if Item 21A = 0003.2 or 22A, 22B or 22C = G10; else = 0 | -0.8977 |
| Comorbidity Condition Group 24 | Seizure Disorders and Convulsions (HCC79) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #79; else = 0 | -0.3754 |
| Comorbidity Condition Group 25 | Cerebral Hemorrhage (HCC99); Ischemic or Unspecified Stroke (HCC100); Cerebrovascular Atherosclerosis, Aneurysm, and Other Disease (HCC102); Hemiplegia/Other Late Effects of CVA: Hemiplegia/Hemiparesis (HCC103); Late Effects of Cerebrovascular Disease Except Paralysis (HCC105) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #99; HCC #100; HCC #102; HCC #103; HCC #105; =0 if Primary Diagnosis Group = Stroke; else = 0 | -1.1349 |
| Comorbidity Condition Group 26 | Atherosclerosis of the Extremities with Ulceration or Gangrene (HCC106) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #106; =0 if Item 21A = 0017.4; else = 0 | -0.8867 |
| Comorbidity Condition Group 27 | Aspiration, Bacterial, and Other Pneumonias: Aspiration and Specified Bacterial Pneumonias (HCC114) | =1 in Item 24 = see Crosswalk ICD-10 codes to HCC #114; =0 if Item 21A = 17.51 or 17.52; else = 0 | -0.2540 |
| Comorbidity Condition Group 28 | Legally Blind (HCC119) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #119; else = 0 | -1.5453 |
| Comorbidity Condition Group 29 | Proliferative Diabetic Retinopathy and Vitreous Hemorrhage (HCC122); Diabetic and Other Vascular Retinopathies (HCC123) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #122; HCC #123; else = 0 | -0.7617 |
| Comorbidity Condition Group 30 | Dialysis and Chronic Kidney Disease - Stage 5: Dialysis Status (HCC134); Chronic Kidney Disease, Stage 5 (HCC136) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #134; HCC #136; =0 if Item 21A = 0017.9 or 22A, 22B or 22C = N18.5; else = 0 | -0.9853 |
| Comorbidity Condition Group 31 | Chronic Kidney Disease, Severe (Stage 4) (HCC137) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #137; =0 if 22A, 22B or 22C = N18.1 or N18.2 or N18.3 or N18.4 or N18.9; else = 0 | -0.3839 |
| Comorbidity Condition Group 32 | Urinary Obstruction and Retention (HCC142) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #142; else = 0 | -0.4780 |
| Comorbidity Condition Group 33 | Chronic Ulcer of Skin, Excluding Pressure Ulcer (HCC161) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #161; =0 if Item 21A = 0017.7; else = 0 | -0.6012 |
| Comorbidity Condition Group 34 | Major Head Injury (HCC167) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #167; =0 if Primary Diagnosis Group = Traumatic Brain Dysfunction; else = 0 | -0.2234 |
| Comorbidity Condition Group 35 | Major Fracture, Except of Skull, Vertebrae, or Hip (HCC171) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #171; =0 if Item 21A = 0008.2 or 0008.4 or 0008.9 or 0014.9; else = 0 | -1.6168 |
| Comorbidity Condition Group 36 | Complication of Specified Implanted Device or Graft (HCC176) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #176; =0 if Primary Diagnosis Code = Hip and Knee Replacements; =0 if Item 21A = 0017.8; else = 0 | -0.7985 |
| Comorbidity Condition Group 37 | Amputations: Traumatic Amputations and Complications (HCC173) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #173; =0 if Primary Diagnosis Group (calculated above) = Amputation; else = 0 | -0.0492 |
| Comorbidity Condition Group 38 | Amputation Status, Lower Limb/Amputation Complications (HCC189) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #189; =0 if Primary Diagnosis Group (calculated above) = Amputation; else = 0 | -0.2744 |
| Comorbidity Condition Group 39 | Amputation Status, Upper Limb (HCC190) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #190; =0 if Primary Diagnosis Group (calculated above) = Amputation; else = 0 | -0.1697 |
| Comorbidity Condition Group 40 | Kidney Transplant Status (HCC132) | =1 if Item 24 = see Crosswalk ICD-10 codes to HCC #132; =0 if Item 21A = 0017.8 or 0017.9; else = 0 | -0.0075 |

**Appendix B:  
Reliability and Validity Testing**

**B.1 Overview of Reliability and Validity Testing**

The goal of reliability testing is to ensure that items on an assessment obtain consistent results when administered or used by different clinicians. Validity testing examines whether an item or scale measures what it is intended to measure. The functional status items underwent reliability testing at the item- and scale-level in multiple types of providers in conjunction with the Post-Acute Care Payment Reform Demonstration. Item-level testing included inter-rater reliability testing within facilities and the use of videotaped standardized patients for inter-rater reliability testing across facilities/care settings. Additional testing focused on the items and scales and included internal consistency, factor analysis, and Rasch analysis. A brief summary of this testing is provided below; full reports describing the testing are available at <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/CARE-Item-Set-and-B-CARE.html> exit icon.

**B.2 Traditional Inter-rater Reliability Study**

The reliability of the functional items was tested in a subset of 34 providers from each of the five levels of care (acute hospitals, HHAs, IRFs, LTCHs, and SNFs) distributed across 11 geographic areas. Each provider completed a duplicate CARE Item Set (admission or discharge assessment) on 15–20 patients included in the Post-Acute Care Payment Reform Demonstration (10–15 patients in the home health setting), in accordance with the guidelines and protocols.

Providers were asked to enroll a convenience sample of a set number of Medicare patients each month, representing a range of function and acuity. The overall patient sample size for each of the functional items was 450 for self-care items and 449 for mobility items (448 for transfers). After exclusions for missing data (unknown/not attempted/inapplicable), the effective sample sizes for the reliability testing were as follows:

* Eating: 401
* Oral hygiene: 414
* Toilet hygiene: 416
* Upper body dressing: 420
* Lower body dressing: 413
* Lying to sitting on the side of the bed: 412
* Sitting to standing: 387
* Chair/bed to chair transfer: 392
* Toilet transfer: 361
* Walk 150 feet: 68
* Walk once standing: 52
* Wheel in room: 46

The inter-rater reliability study included patients who were assessed by two different clinicians (raters), and the agreement of the clinicians’ rating was calculated. Clinicians were instructed to have pairs of raters complete both patient assessments at the same time. Responses to items were obtained by direct observation of the patient by the clinician, and occasionally, supplemented by one or more of the following predetermined, matched methods: patient interviews (with each team member taking turns conducting and observing patient interviews); interviews with relatives/caregivers of the patient for certain items; and/or interviews with staff caring for the patient and/or chart review. Rater pairs were instructed to determine in advance which methods would be used to score the particular CARE items and to have both raters use the same methods. Raters were encouraged to divide hands-on assistance to the patient as evenly as possible for items that required hands-on assistance. Raters were instructed not to discuss item scoring during the assessment, nor to share item scores until the data were entered into the study database and finalized. Providers submitted data via the online CARE application for both assessments in each pair.

For categorical items, kappa statistics (kappa) 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 International calculated only 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. 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.[[7]](#footnote-7),[[8]](#footnote-8) 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 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 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 core self-care items show similar agreement to the overall estimates. The lower-body dressing item had the highest overall weighted kappa (0.855), whereas the eating item had the lowest (0.798). Unweighted overall kappas ranged from 0.636 (toileting) to 0.598 (oral hygiene). Acute hospitals had the highest weighted kappas across all self-care items.

The weighted kappa values for the mobility items ranged between 0.558 for walk 150 feet to 0.901 for sitting to standing and chair/bed to chair transfer. Unweighted kappas ranged from 0.667 for walk once standing to 0.762 for sit to stand. Provider-specific analyses of core mobility items show similar agreement to the overall estimates. The sit-to-stand and chair transfer items both had a weighted kappa of 0.901, whereas the lying to sitting item had a weighted kappa of 0.855. Unweighted overall kappas ranged from 0.693 (lying to sitting) to 0.762 (sitting to standing).

**B.3 Videotaped Standardized Patients Reliability Study**

For the video reliability study, which was designed to examine the level of clinician agreement across care settings, clinicians in each setting 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. The mid-level functional activities were considered to be the most challenging for clinicians to score and are thus of particular interest in establishing reliability. Each clinician involved in the video study watched three videos and assessed the patients according to the study guidelines and protocols. Each video was approximately 20 minutes long and had a corresponding item set arranged in the sequence in which the items appeared in the video.

The sample included 28 providers (550 assessments), which included 3 acute hospitals (15 assessments [3%]); 9 HHAs (118 assessments [22%]); 8 IRFs (237 assessments [43%]); 3 LTCHs (114 assessments [21%]); and 5 SNFs (66 assessments [12%]). Participating providers included case managers (6% of assessments), occupational therapists (14% of assessments), physical therapists (21% of assessments), registered nurses (47% of assessments), speech therapists (5% of assessments), and others, mostly licensed practical nurses (LPNs; 8% of assessments).

Two main analytic approaches were used for assessing the video reliability of the CARE items, adhering closely to the methods used by Fricke et al.[[9]](#footnote-9) in their video reliability study of the FIM®[[10]](#footnote-10) instrument. First, percent agreement with the mode response was calculated for each CARE item 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.

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 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 video reliability study indicated substantial agreement with the mode and clinical team for the lying-to-sitting, sit-to-stand, chair/bed to chair transfer, and toilet transfer items (greater than 76%). Although rates of agreement with the mode and clinical team response were generally identical, for the toilet transfer item, the clinical team agreement is slightly lower. The items for walking and wheeling distances showed more variable levels of agreement across disciplines, with overall agreement generally in the moderate range (50–78%). For the Walk in Room item, there was a notable decrease in the agreement with the clinical team compared to agreement with the mode. This occurred because in two of the four videos where this item was assessed, the clinical team response differed from the mode.

**B.4 Scale-level Reliability Results: 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. General 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.[[11]](#footnote-11)

Assessments of individual self-care and mobility subscales at both admission and discharge tend to show good reliability statistics (Cronbach’s Alpha of at least 0.80) within their specified subscales. Reliability estimates by provider type show 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.

The following table shows the findings from the Cronbach’s alpha internal consistency evaluation mentioned above.

**Table B-1  
CARE functional status internal consistency reliability summary by provider type**

| CARE analytic set | Overall alpha | HHA alpha | SNF alpha | IRF alpha | LTCH alpha |
| --- | --- | --- | --- | --- | --- |
| Self-Care | 0.96 | 0.94 | 0.95 | 0.95 | 0.96 |
| Mobility | 0.96 | 0.94 | 0.95 | 0.96 | 0.97 |

**B.5 Scale-level Reliability and Validity Testing: Rasch Analysis**

Because we are measuring a latent trait—a concept that is not measured directly, but that relies on activities that can be directly observed—we used the one-parameter Rasch model to gain a better understanding of the functional status activities. More specifically, we examined the order of functional status items (from least challenging to most challenging) that characterize the concepts of the self-care and mobility.

Rasch analysis uses the scores from the functional assessment items to create the equivalent of a functional status “ruler” (i.e., scale). Rasch analysis uses the available data to estimate a person’s location along the “ruler;” therefore, analyses can be conducted if some data are missing. Rasch analysis can also inform the optimal selection of key items in order to construct functional status scales that sufficiently span an entire range of patient functioning, so that both the least able and most able (lowest- and highest-functioning) patients are adequately measured. In addition, Rasch analysis can indicate where items overlap or are redundant in terms of the level of function they capture.

Rasch analysis has been used to examine the FIM® instrument,[[12]](#footnote-12),[[13]](#footnote-13),[[14]](#footnote-14),[[15]](#footnote-15) the Minimum Data Set (MDS),[[16]](#footnote-16) and the Outcome and Assessment Information Set (OASIS).[[17]](#footnote-17) Rasch analysis has also been used to examine the extent to which existing functional assessment instruments (e.g., the FIM® instrument, MDS 2.0) capture the same construct.[[18]](#footnote-18)

Rasch measurement is based on a probabilistic model that describes the association between a person’s underlying ability level and probability of a particular item response, and summarizes a patient’s position along a “ruler” that represents a latent trait or concept (e.g., self-care or mobility).[[19]](#footnote-19) In essence, the Rasch analysis creates a ruler based on the domain measured (e.g., mobility) that can be used to assess the abilities of the patients. The analysis also provides information on the hierarchy of item difficulty (from easy to hard) that can be used to evaluate the construct validity of a set of items. In addition, the Rasch analysis provides information about the level of challenge associated with each item rating scale (“dependent” through “independent”). For example, an item with a low difficulty estimate (e.g., eating) would be more likely to be completed with little or no help by patient’s items that are more challenging (e.g., 12 steps), where most patients would find completing this activity challenging. Finally, the Rasch analysis can provide information on items that do not fit into the single theorized concept through “item misfit” statistics, which may indicate that the item needs further evaluation before it is included on future administrations of the subscale. The infit mean square is an indicator of the degree to which patient responses are similar to what would be expected (i.e., predicted) by the measurement model. The acceptable range is generally 0.6 to 1.4. If the item values are above this range, it reflects that person response patterns are erratic, generally suggesting that the item is not measuring the same construct as other items. Infit mean squares above 1.4 are considered to be unacceptably unexpected[[20]](#footnote-20) and indicate that the item most likely does not reflect the same construct as the other items included in the scale; for example, a need for assistance with self-care.

RTI used Rasch analysis to examine the extent to which the items worked together to define a coherent concept. This was conducted separately for the self-care and mobility items. Item fit statistics were examined as an indication of how well all items work together to describe the overall construct (self-care or mobility). The Rasch analysis provides insight into how the items work together as a subscale, including the hierarchy of item difficulty (ordering from easy to difficult) and item fit to the model.

Examinations of these Rasch analysis results reveal that the mobility and self-care item hierarchies make sense clinically and that the operational definitions of the constructs maintain general stability from admission to discharge. Some items have fit statistics outside the acceptable range (e.g., pick up object from floor), but members of the Technical Expert Panel noted that this is an important assessment given the risk of falls.

RTI examined how well the items selected measure the persons in the data set for both self-care and mobility items. RTI examined the extent to which person response patterns fit the assumptions of the measurement model using the same range of infit statistics identified above. RTI examined the extent to which persons are effectively measured (ceiling and floor effects) in each setting overall and for admission and discharge time points. The mobility and self-care items were found to be well targeted to the range of patient ability sampled within this post-acute care population.

RTI established that the six steps of the CARE rating scale are operating as intended, both overall and for individual items on the self-care and mobility subscales. The probability that a person will be scored on a particular rating scale step varies depending on the functional ability of the person. That is, very able people will be more likely to be scored as ‘5’ and ‘6’ than as ‘1’ and ‘2.’ Looking empirically at these distributions, one should see the transitions from one step to the next (called thresholds) proceed monotonically and distinctly across the range of person abilities. In other words, there should always be some point along the range at which each rating-scale step is more probable than another step. When a rating-scale step is not more probable at any point, it suggests that raters are not able to use that step to consistently distinguish patient ability at that level.

1. <https://www.huduser.gov/portal/datasets/usps_crosswalk.html> [↑](#footnote-ref-1)
2. <https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.aspx> [↑](#footnote-ref-2)
3. [↑](#footnote-ref-3)
4. <https://www.huduser.gov/portal/datasets/usps_crosswalk.html> [↑](#footnote-ref-4)
5. <https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.aspx> [↑](#footnote-ref-5)
6. [↑](#footnote-ref-6)
7. Hirdes JP, Smith TF, Rabinowitz T, et al. The Resident Assessment Instrument-Mental Health (RAI-MH): inter-rater reliability and convergent validity. *J Behav Health Serv Res.* 29(4):419-432, 2002 [↑](#footnote-ref-7)
8. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. *Oxford University Press*, 1995. [↑](#footnote-ref-8)
9. Fricke J, Unsworth C, Worrell D. Reliability of the Functional Independence Measure with Occupational Therapists. *Australian Occupational Therapy Journal* 40(1):7-15, 1993. [↑](#footnote-ref-9)
10. FIM® is a trademark of Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. [↑](#footnote-ref-10)
11. Aron A, Aron EN *Statistics for Psychology*. 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1999. [↑](#footnote-ref-11)
12. Granger CV, Hamilton BB, Linacre JM, et al. Performance profiles of the functional independence measure. *Am J Phys Med Rehabil.* 72(2):84-89, 1993. [↑](#footnote-ref-12)
13. Linacre JM, Heinemann AW, Wright BD, et al. The structure and stability of the Functional Independence Measure. Archives of Physical Medicine & Rehabilitation.75(2):127-132, 1994 [↑](#footnote-ref-13)
14. Wright BD, Linacre JM, Smith RM, et al. FIM measurement properties and Rasch model details. Scandinavian Journal of Rehabilitation Medicine, 29(4):267-272, Dec. 1997. [↑](#footnote-ref-14)
15. Heinemann AW, Linacre JM, Wright BD, et al. Relationships between impairment and physical disability as measured by the functional independence measure. Arch Phys Med Rehabil. 74(6):566-573, 1993. [↑](#footnote-ref-15)
16. Wang YC, Byers KL, Velozo CA. Rasch analysis of Minimum Data Set mandated in skilled nursing facilities. J Rehabil Res Dev. 45(9):1385-1399, 2008. [↑](#footnote-ref-16)
17. Fortinsky RH, Garcia RI, Joseph Sheehan T, et al. Measuring disability in Medicare home care patients: application of Rasch modeling to the outcome and assessment information set. Med Care. 41(5):601-615, 2001. [↑](#footnote-ref-17)
18. Velozo CA, Byers KL, Wang YC, et al. Translating measures across the continuum of care: using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set. J Rehabil Res Dev. 44(3):467-478, 2007. [↑](#footnote-ref-18)
19. Wright BD, Stone MH. Best Test Design. Rasch Measurement. 1979. [↑](#footnote-ref-19)
20. Wright BD, Linacre JM, Gustafson J, et al. Reasonable mean-square fit values. Rasch Measurement Transactions. 8(3):370, 1994. [↑](#footnote-ref-20)