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

**Measure Number** (*if previously endorsed*)**:** 0677

**Measure Title**: Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)

**Date of Submission**: 4/16/2018

**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 (including 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 decision-making) 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: Nursing Home Minimum Data Set (MDS) 3.0. | other: Nursing Home Minimum Data Set (MDS) 3.0. |

**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 data set used for testing was the Nursing Home Minimum Data Set (MDS) 3.0.

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

Two studies were used in the testing of this measure; they are described in greater detail below.

1. RAND Development and Validation of MDS 3.0 study: August 2006 to February 2007 (Saliba & Buchanan, 2008).
2. RTI International Analysis of MDS 3.0 data: Quarter 1, 2017 and Quarter 2, 2017 (in addition, performance scores were presented for Quarter 1, 2011 – Quarter 2, 2017 in Section 2b1).

Saliba, D., & Buchanan, J. (2008, April). *Development and validation of a revised nursing home assessment tool: MDS 3.0*. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation. Retrieved from <http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf>.

**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*)

1. The RAND Development and Validation of MDS 3.0 study sample included a representative sample of for-profit and not-for-profit facilities, and hospital-based and freestanding facilities, which were recruited for the study. It included 71 community nursing facilities in 8 states and 19 Veterans Affairs (VA) nursing homes. This study tested item-level reliability and validity of the items used to calculate the pain measures by comparing item coding among gold-standard nurses (349 cases), and comparing item coding between gold-standard nurses and participating facility staff (900 cases) (Saliba & Buchanan, 2008).
2. RTI facility-level analyses of MDS 3.0 data include all facilities with sufficient sample size (*n* ≥ 20) to publicly report this measure in Quarter 2, 2017 (*k* = 13,691).

Saliba, D., & Buchanan, J. (2008, April). *Development and validation of a revised nursing home assessment tool: MDS 3.0*. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation. Retrieved from <http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf>.

SOURCE: RTI analysis of MDS 3.0 data for Quarter 2 2017 (programming reference: rn01/ac362\_request\_q2627\_677.log)

**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*)

1. The RAND Development and Validation of MDS 3.0 study sample included 3,822 residents from community nursing homes and 764 residents from VHA nursing homes (Saliba & Buchanan, 2008).
2. The sample for the RTI analysis of MDS 3.0 includes all long-stay residents that meet the denominator inclusion criteria for this measure in facilities with sufficient sample size (*n* ≥ 20, *k* = 13,691) to report this measure (*n* = 813,872) in Quarter 2, 2017.

SOURCE: RTI analysis of MDS 3.0 data for Quarter 2 2017 (programming reference: rn01/ac362\_request\_q2627\_677.log)

***Table 1*** below presents the characteristics of long-stay residents counted in the denominator for this measure in Quarter 2, 2017 before applying facility sample size restrictions and without excluding those residents without both a prior and target assessment (*n* = 838,842); the *n* for each resident characteristic varies due to the proportion of missing data for that characteristic.

Table 1.   
Characteristics of Long-Stay Residents Included in Analyses of NQF #0677 (Quarter 2, 2017)

| Resident characteristics | Frequency (n) | Percentage (%) |
| --- | --- | --- |
| Sex |  |  |
| Female | 544,102 | 64.9% |
| Male | 294,740 | 35.1% |
| Race |  |  |
| White | 635,209 | 75.7% |
| Non-white | 203,151 | 24.2% |
| Missing | 482 | <1.0% |
| Age |  |  |
| <65 | 133,213 | 15.9% |
| 65-74 | 150,382 | 17.9% |
| 75-84 | 214,025 | 25.5% |
| 85+ | 341,222 | 40.7% |
| Diagnoses |  |  |
| Arthritis | 258,001 | 30.9% |
| Osteoporosis | 116,310 | 13.9% |
| Hip Fracture | 13,577 | 1.7% |
| Other Fracture | 24,098 | 3.0% |
| Depression | 433,082 | 53.4% |
| Stroke | 98,284 | 12.1% |
| Alzheimer's Disease | 107,035 | 13.2% |
| Non-Alzheimer's Dementia | 371,583 | 45.9% |
| Malnutrition or at risk for malnutrition | 31,055 | 3.7% |
| Cancer | 53,214 | 6.4% |
| Anemia | 245,866 | 30.3% |
| Heart Failure | 175,942 | 21.7% |
| Hypertension | 634,161 | 78.3% |
| Diabetes Mellitus | 294,783 | 35.1% |
| Anxiety Disorder | 265,825 | 31.7% |
| Asthma, Chronic Obstructive Pulmonary Disease, or  Chronic Lung Disease | 193,642 | 23.9% |

Analysis date: 12/6/17

SOURCE: RTI analysis of MDS 3.0 data for Quarter 2, 2017 (programming reference: av010/nb008\_request\_v1\_q\_26\_27\_677.log)

Saliba, D., & Buchanan, J. (2008, April). *Development and validation of a revised nursing home assessment tool: MDS 3.0*. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation. Retrieved from <http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf>.

**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**.

The analyses used the same data as described above in sections 1.5 and 1.6.

**Data for Critical Data Elements**

RAND reliability analysis of data elements used the same sample as described in sections 1.5 and 1.6 (Saliba & Buchanan, 2008).

**Data for Measure Performance Score Testing**

RTI analyses used the same data as described in sections 1.5 and 1.6.

Saliba, D., & Buchanan, J. (2008, April). *Development and validation of a revised nursing home assessment tool: MDS 3.0*. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation. Retrieved from <http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf>.

**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.

Analyses are based on social risk factor variables related to self-reported pain and available in the MDS 3.0, including Medicaid status, gender, marital status, and age.

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

**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*)

**Critical Data Element Reliability**

1. The national test of MDS 3.0 items examined the agreement between assessors (reliability). Quality Improvement Organizations were employed to identify gold-standard (research) nurses and recruit community nursing facilities to participate in the national evaluation (Saliba & Buchanan, 2008). The gold-standard nurses were trained in the MDS 3.0 instrument, and they, in turn, trained a facility nurse from each participating nursing facility in their home states. Residents participating in the test were selected to capture a representative sample of short- and long-stay residents. In this national test, the agreement between gold-standard nurses and between gold-standard and facility nurses on pain treatment regimen items, pain interview items, and staff observed pain behaviors were examined. For the pain interview items used to create this measure, reliability was assessed for the self-reported items as reported to gold-standard and facility nurses. In addition, the temporal reliability of the pain interview items was assessed by administering the assessment twice within a 24-hour period; a 24-hour period was used to prevent actual changes in resident pain from influencing results of the reliability analyses. Agreement between gold-standard nurses and between gold-standard and facility nurses was also examined for MDS Section C items, which comprise the Brief Interview for Mental Status (BIMS), a structured cognitive assessment; items for Section C are used to develop the risk adjustment covariate, described in 2b3.1.1. Cohen’s kappas were calculated to assess item reliability. Kappa is a statistical measure of inter-rater agreement for qualitative data, ranging from 0.0 to 1.0, where a rating of greater than 0.60 is considered substantial agreement (Landis & Koch, 1977).

In recent years, patient-reported outcomes and patient satisfaction with care have been emphasized in research and policy initiatives. In particular, patient interview regarding the presence and severity of pain is considered to be the most reliable and accurate approach to pain assessment (American Geriatrics Society, 2002; Department of Veterans Affairs, 2003); however, it is still important to examine the agreement between assessors to ensure that there is not unintended variation in provider interpretation of patient-reported information.

Saliba, D., & Buchanan, J. (2008, April). *Development and validation of a revised nursing home assessment tool: MDS 3.0*. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation. Retrieved from <http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf>.

Landis, JR, Koch, GG. The measurement of observer agreement for categorical data. *Biometrics* *33*(1), p 159-174, 1977.

American Geriatrics Society Panel on Persistent Pain in Older Persons. The management of persistent pain in older persons. *Journal of the American Geriatrics Society 50*, p S205-244, 2002.

Department of Veterans Affairs. VHA directive 2003-021: pain management. 2003.

**Performance Measure Score Reliability**

2.a. Signal to noise analysis: If a measure is reliable, then true differences in provider performance should explain a substantial proportion of the variance in quality measure scores. A signal-to-noise analysis was performed to determine what proportion of total variance in the measure is attributable to differences among providers. This analysis used logistic regression of the measure numerator triggering for Quarter 2, 2017. We ran a logistic regression analysis with one term (a binary variable equal to 1 if the measure numerator is triggered and 0 if otherwise; please refer to S.4 and S.5 for more details on the measure numerator specifications) with facility random effects to obtain an estimate of ρ, the proportion of the total variance contributed by the facility-level variance component (i.e.., *ρ* = ).

2.b. Split-Half reliability analysis: Split-half reliability assesses the internal consistency of a quality measure by randomly dividing the residents within each nursing facility into two halves and calculating the correlation between the nursing facility’s quality measure scores on the basis of the two randomly divided halves. When a nursing facility’s residents, randomly divided, have similar scores to one another, the quality measure score is more likely to reflect systematic differences in nursing home-level quality rather than random variation. In this analysis, we conducted a split-half reliability analysis on all facilities with 20 or more residents counted in the measure denominator. We used the Pearson Product-Moment Correlation r, Spearman Rank Correlation ρ, and Intraclass Correlation Coefficient (ICC) to measure the internal reliability.

**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*)

**Critical Data Element Reliability**

1. For the pain items, the average kappa for gold-standard nurse to gold-standard nurse agreement was 0.961, and the average kappa for gold-standard nurse to facility nurse agreement was 0.967. The average kappa for temporal reliability was 0.920. All three kappas are well above the value of 0.60, which is generally considered substantial agreement (Saliba & Buchanan, 2008).

For the cognition items, the average kappa for gold-standard nurse to gold-standard nurse agreement was 0.977, and the average kappa for gold-standard nurse to facility nurse agreement was 0.973, demonstrating very strong agreement between assessors on these items.

Saliba, D., & Buchanan, J. (2008, April). *Development and validation of a revised nursing home assessment tool: MDS 3.0*. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation. Retrieved from <http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf>.

**Performance Measure Score Reliability**

2.a. Signal to Noise: The signal-to-noise ratio for this measure was high at 0.289 (p < 0.001) indicating that 28.9% of the variance in scores for this measure in Quarter 2, 2017 was explained by facility characteristics (including the underlying quality of care in each facility). Thus, this measure is reliable in separating facility characteristics from the noise of population variance.

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: rn01\db363\_request\_q2627\_677.log)

2.b. Split-Half Reliability Analysis: The split-half correlations were strong (r = 0.68, ρ = 0.66, p < .001), and the ICC was 0.68 (p < .001), demonstrating strong internal reliability.

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: av12\av12\_request\_q2627\_677.log)

**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?*)

**Critical Data Element Reliability**

The RAND Development and Validation of MDS 3.0 national pilot test study demonstrated good reliability for MDS 3.0 items used to calculate this measure.

**Performance Measure Score Reliability**

RTI’s analyses show that the measure is largely stable, is able to make meaningful dinstinctions between different facilites’ quality on this measure’s intended dimension, the measure has strong internal reliability, and the differences in the facilities can explain about a quarter of the variance in this measure. In sum, this measure is reliable.

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

**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)*

**Critical Data Element Validity**

1. RAND Development and Validation of MDS 3.0: The RAND validation of MDS 3.0 tested the criterion validity of the items used to calculate this measure and of the items used to calculate the risk adjustment covariate by comparing how different nurses assessed the same residents using MDS 3.0. They compared gold-standard nurses to staff nurses trained by the gold-standard nurses. The Kappa statistic was calculated (Saliba & Buchanan, 2008).

In addition, the RAND validation of the MDS 3.0 tested the correlation between the BIMS and the Modified Mental State Exam (3MS), which was used as the gold-standard measure of cognitive function. The 3MS is an expanded version of the Mini Mental State Exam (MMSE), and research demonstrates that the 3MS has greater reliability and validity than the MMSE (Saliba & Buchanan, 2008; Bland & Newman, 2001; Grace, Nadler, White et al, 1995; McDowell, Kristjansson, Hill et al, 1997; Teng & Chui, 1987).

For additional analyses of item-level validity, please refer to Appendix C, which contains analyses submitted to the NQF Scientific Methods Panel on March 30, 2018.

**Performance Measure Score Validity**

2.a. Correlation with related quality measures: To assess convergent validity, RTI examined whether a facility’s percentile rank on one quality measure in a measure group was correlated with its percentile rank on another quality measure in the same clinically-related group. Specifically, we examined whether a facility’s percentile rank on this measure (NQF# 0677) was correlated with that facility’s performance on the related short-stay pain quality measure (NQF #0676). We hypothesize that a nursing facility’s percentile rank on NQF #0677 and its percentile rank on NQF #0676 should have a strong and positive correlation. Although the short stay and long stay populations that a nursing facility serves likely have different levels of resident acuity, a nursing facility’s ability to manage resident pain should not vary substantially across these two populations.

2.b. Seasonality: Another potential threat to the validity of a quality measure is seasonal variation. If a quality measure score varies substantially from quarter to quarter in a consistent pattern over time corresponding to changes in seasons, it is possible that the validity of the measure is being compromised due to influences not within a nursing home’s control. To address whether seasonal variation might play a role, we examined the trend in the national mean and median for this quality measure score between Quarter 1, 2011 and Quarter 2, 2017.

2.c. Stability analysis: We examined the extent to which relative facility rank changed on this quality measure from Quarter 1 to Quarter 2, 2017. We did this in two ways: 1) RTI evaluated the percentage of facilities that changed in terms of their standardized quality measure score by less than 1 standard deviation, between 1 and 2 standard deviations, between 2 and 3 standard deviations, and 3 standard deviations or more from quarter to quarter; 2) RTI evaluated the percentage of facilities that changed in their percentile ranking (i.e., relative quality measure score) within 1 decile, between 1 and 2 deciles, between 2 and 3 deciles, and 3 or more deciles. Dramatic changes in the quality measure score or facility rank based on the score over time may indicate measure instability, rather than true changes in quality.

2.d. Confidence interval analysis: We examined proportions of facilities with scores for this measure that are significantly different from national facility-level mean, stratified by facility denominator size. A reliable measure should have a high proportion of facilities with scores significantly different than the mean. For this analysis, statistical significance was determined using 95% confidence intervals: a facility’s quality measure score was significantly different from the national mean if the national mean was not included in the facility’s 95% confidence interval. Because this measure is focusing on an undesirable outcome, high-performing facilities should have scores that are significantly below average, and scores of low-performing facilities should be significantly above average. We stratified the analysis by facility denominator size to examine whether this feature of the measure varies by size.

For additional analyses of performance measure score validity, please refer to Appendix C, which contains analyses submitted to the NQF Scientific Methods Panel on March 30, 2018.

Saliba, D., & Buchanan, J. (2008, April). *Development and validation of a revised nursing home assessment tool: MDS 3.0*. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation. Retrieved from <http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf>.

Bland, R. & Newman, S. Mild dementia or cognitive impairment: the modified mini-mental state examination (3MS) as a screen for dementia. *Can J Psychiatry 46*, 506-510. 2001.

Grace, J., Nadler, J., White, D., et al. Folstein vs modified mini-mental state examination in geriatric stroke, stability, validity, and screening utility. *Archives of Neurology 52*, 477-484. 1995.

McDowell, I., Kristjansson, B., Hill, G., et al. Community screening for dementia: the mini mental state exam (MMSE) and modified mini-mental state exam (3MS) compared. *J Clin Epidemiol 50,* 377-383. 1997.

Teng, E., & Chui, H. The modified mini-mental state (3MS) examination. *J Clin Psychiatry 48*, 314-318. 1987.

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

**Critical Data Element Validity**

1. The RAND national pilot test of the MDS 3.0 items conducted by Saliba and Buchanan indicated that the pain items and the cognition items are reliable and valid. For the pain items, the average kappa for gold-standard nurse to gold-standard nurse agreement was 0.961, and the average kappa for gold-standard nurse to facility nurse agreement was 0.967, both well above the value of 0.60, which is generally considered substantial agreement. For the cognition items, the average kappa for gold-standard nurse to gold-standard nurse agreement was 0.977, and the average kappa for gold-standard nurse to facility nurse agreement was 0.973, demonstrating very strong agreement between assessors on these items.

The correlation between the BIMS and the 3MS was very strong ( = 0.906) and statistically significant (p < 0.001).

**Performance Measure Score Validity**

2.a. Correlation with related quality measures: Among facilities who could report both measures, RTI calculated the correlation between the facility’s percentile rank on NQF #0676 (Percent of Residents Who Self-Report Moderate to Severe Pain (Short Stay)) and the facility’s percentile rank on NQF #0677 (Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)) and found a strong ( = 0.655) and statistically significant (p < 0.001) direct correlation.

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: rn06\ac359\_request q2627\_677.log)

2.b. Seasonality: RTI examined the national-level mean and median quality measure scores for each quarter from Quarter 1, 2011, to Quarter 2, 2017. The results are presented in ***Figure 1***. The national-level means and medians have both decreased almost monotonically since Quarter 1 of 2011. These results show no evidence of seasonal variation. Further, this also indicates that facilities may have improved in pain management over this period.

Figure 1  
Seasonal (Quarterly) Variation, NQF #0677 Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)

SOURCE: RTI analysis of MDS 3.0 episode files for Quarter 1, 2011–Quarter 2, 2017

2.c. Stability analysis: ***Figure 2*** illustrates the changes in facility rank by quality measure score from Quarter 1, 2017, to Quarter 2, 2017. The majority (60.1%) of facilities ranked within the same decile in both quarters. Shifts of more than 3 deciles were much less common, occurring for approximately 11% of facilities. ***Figure 3*** illustrates change in facility-level scores relative to the overall distribution of scores. Standardized scores for this measure showed similar stability: 77.2% of facilities saw their scores change by less than one standard deviation (i.e., the *z*-score changed by one or less), and only 1.6% of facilities had score changes of more than three standard deviations from one quarter to the next. Thus, both facility scores and relative ranks for this measure are stable from one quarter to the next.

Figure 2  
Decile Change in Facility Ranking from Quarter 1, 2017, to Quarter 2, 2017, NQF #0677 Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)

Analysis date: 4/5/2018

SOURCE: RTI analysis of Q1 2017 and Q2 2017 MDS 3.0 data (programming reference: rn01/ac358\_request\_q2627\_677.log)

Figure 3  
Standardized Score Change from Quarter 1, 2017, to Quarter 2, 2017, NQF #0677 Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)

Analysis date: 4/5/2018

SOURCE: RTI analysis of Q1 2017 and Q2 2017 MDS 3.0 data (programming reference: [rn01\ac355\_request\_q2627\_677.log](file:///X:\004TES~1\Pgm\ykaganova\rn\rn01\db355_request_q2627_677.log))

2.d. Confidence interval analysis: Another measure of consistency is performance relative to the mean: high-performing facilities should have scores that are significantly below-average, and low-performing facilities should be significantly above-average. ***Table 2*** shows the proportions of facilities that scored significantly higher or lower than the national facility-level mean in Quarter 2, 2017. For this analysis, statistical significance was determined using 95% confidence intervals: a facility’s quality measure score was statistically significantly different from the national mean if the national mean was not within that facility’s 95% confidence interval. This analysis was also stratified by decile of facility size.

In general, there were fewer facilities with quality measure scores that were statistically significantly (*p* ≤ .05) higher than the national mean of 7.2% (9.7%), than those with scores that were statistically significantly lower than the national mean (36.5%), indicating that more facilities perform better (lower scores are better) than the national facility-level mean.

The proportions of facilities with scores that are significantly different from the national mean vary as a function of the number of residents included in the denominator for this measure; the percentage of facilities which have scores that are statistically significantly different from the mean increase with the number of residents. Increases in the facility-level sample size lead to reductions in the standard error of facility-level scores. Thus, it appears that changes in the reliability of this measure for the larger facilities are due to the greater statistical reliability that accompanies increased sample size.

Overall, almost half (46.2%) of facilities were significantly different from the national mean in Quarter 2, 2017, indicating that there are meaningful differences in facility-level scores for this measure.

Table 2  
Proportion of Facilities with Scores Significantly Different from the National Facility-Level Mean, Stratified by Facility Denominator Size for NQF #0677 Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)

| Decile of denominator size in residents | *k* | | | Number of facilities with 95% confidence interval lower than national mean (%) | | Number of facilities with 95% confidence interval higher than national mean (%) | | Total number of facilities with scores significantly different from mean (%) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1st Decile (*n*= 20 to 27) | | 1,419 | 441 (31.1%) | | 100 (7.0%) | | 541 (38.1%) | |
| 2nd Decile (*n* = 28 to 33) | | 1,352 | 378 (28.0%) | | 107 (7.9%) | | 485 (35.9%) | |
| 3rd Decile (*n* = 34 to 39) | | 1,383 | 381 (27.5%) | | 123 (8.9%) | | 504 (36.4%) | |
| 4th Decile (*n* = 40 to 46) | | 1,518 | 504 (33.2%) | | 156 (10.3%) | | 660 (43.5%) | |
| 5th Decile (*n* = 47 to 52) | | 1,289 | 446 (34.6%) | | 125 (9.7%) | | 571 (44.3%) | |
| 6th Decile (*n* = 53 to 59) | | 1,380 | 489 (35.4%) | | 144 (10.4%) | | 633 (45.9%) | |
| 7th Decile (*n* = 60 to 67) | | 1,380 | 489 (35.4%) | | 159 (11.5%) | | 648 (47.0%) | |
| 8th Decile (*n* = 68 to 77) | | 1,248 | 513 (41.1%) | | 117 (9.4%) | | 630 (50.5%) | |
| 9th Decile (*n* = 78 to 97) | | 1,367 | 631 (46.2%) | | 144 (10.5%) | | 775 (56.7%) | |
| 10th Decile (*n* = 98 to 552) | | 1,355 | 725 (53.5%) | | 155 (11.4%) | | 880 (64.9%) | |
| Total | | 13,691 | 4,997 (36.5%) | | 1,330 (9.7%) | | 6,327 (46.2%) | |

NOTE: *k* = number of facilities that meet minimum requirements for public reporting this quality measure.

Analysis date: 4/5/2018

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: rn01\ac364\_request\_q2627\_677.log)

**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?*)

**Critical Data Element Validity**

The items used to define self-reported pain and the risk adjustment covariate have good item level criterion validity (based on the results of the national pilot test of the MDS 3.0).

**Performance Measure Score Validity**

RTI’s analyses indicated that this measure is a valid measurement of self-reported pain. Facilities’ scores on the long-stay pain measure (NQF #0677) are strongly correlated with their scores on the measure of self-reported pain among short-stay residents (NQF #0676), demonstrating convergent validity.

This measure does not vary substantially from quarter to quarter corresponding to changes in seasons; thus, seasonality is not a threat to validity for this measure. This measure is also stable over time and demonstrates substantial variation.

Please see section 2b6 for analysis of the impact of missing data on this measure, which also speaks to validity.

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

**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*)

The denominator exclusion criteria for this quality measure are as follows: 1) The target assessment is an admission assessment, a PPS 5-day assessment or a PPS readmission/return assessment; 2) The resident cannot self-report; 3) The target assessment indicates that the resident had pain or hurting at any time in the last 5 days, but the pain intensity item indicates no pain; or 4) There are missing data in the responses to the relevant pain assessment items in the target assessment.

RTI examined the frequency and proportion of residents excluded from this measure for each of the exclusion criteria for this quality measure.

Exclusion criterion 4 relating to missing data is assessed in greater detail in section 2b6.

**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 281,972 residents were excluded from this quality measure based on the measure denominator exclusions as described above (i.e., 25.2% of long-stay residents in Quarter 2, 2017); please note that exclusion criteria are not mutually exclusive.

173,982 (15.5% of long-stay residents in Quarter 2, 2017) residents were excluded because they were unable to participate in the pain assessment interview (i.e., they were rarely or never understood (J0200 = [0]) or they were unable to answer when asked, “Have you had pain or hurting at any time in the last 5 days? (J0300 = [9])). While this is a substantial proportion of the long-stay resident population, there have not been major changes over time to the proportion excluded for this reason. This measure includes only those individuals who can self-report pain; those unable to self-report would have their pain assessed by nursing facility staff assessment. The decision to focus on self-reported pain was previously well vetted with the Nursing Home Quality Measure (NHQM) project TEP and based on literature showing results from self-report and staff assessment are not comparable.

534 residents (<0.1% of long-stay residents in Quarter 2, 2017) were excluded because data on pain items were inconsistent (item J0300 indicated pain or hurting but item J0600A indicated a pain rating of [0]).

96,793 residents (8.6% of long-stay residents in Quarter 2, 2017) were excluded due to missing data on pain items. Of the 96,793 residents excluded due to missing data on pain items, 62,942 (65.0%) had target assessments that were unplanned discharge assessments. If the unplanned discharge item set is used, the resident interview items are not present and there is no opportunity to conduct a resident interview or staff assessment. In these cases, the MDS 3.0 RAI manual allows for the entry of dashes in lieu of pain data. This is, therefore, a logical reason to exclude those residents for which no data are available. A possible alternative was to use data from the assessment prior to the unplanned discharge assessment. However, using data from the assessment prior to the unplanned discharge would be potentially problematic, given that the assessment before the unplanned discharge for long-stay residents might be from as much as 90 days prior to discharge.

14,514 residents (1.3% of long-stay residents in Quarter 2, 2017) were excluded because their target assessments were either admission, 5-day, or readmission/return assessments.

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: rn01\db360\_request\_q2627\_677.log; ljc04\ljc04\_request\_677; rn16\rn16\_request\_q2627\_677.log; av\db361\_rerun.log)

**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)

Most denominator exclusions for this measure occur because of missing data; impact of missing data on this quality measure is presented in detail in section 2b6.

15.5% of long-stay residents were excluded because they were unable to participate in the pain assessment interview (i.e., they were rarely or never understood (J0200 = [0]) or they were unable to answer when asked, “Have you had pain or hurting at any time in the last 5 days? (J0300 = [9])). While this is a substantial proportion of the long-stay resident population, there have not been major changes over time to the proportion excluded for this reason. This measure includes only those individuals who can self-report pain; those unable to self-report would have their pain assessed by nursing facility staff assessment. The decision to focus on self-report was previously well vetted with the Nursing Home Quality Measure (NHQM) project TEP and based on literature showing that results from self-report and staff assessment are not comparable.

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: rn16\rn16\_request\_q2627\_677.log)

**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** one **risk factor**

**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.**

This measure is risk-adjusted for cognitive independence or modified independence in daily decision-making using a logistic regression. The measure is calculated as follows:

Step 1: Identify the total number of long-stay residents who do not meet the exclusion criteria, with a selected target assessment (OBRA, PPS, or discharge) during the quarter.  
  
Step 2: Calculate the facility-level observed score (steps 2a through 2b below).

Step 2a: Starting with the set of residents identified in Step 1, determine the total number of long-stay residents with a selected target assessment that meets the numerator inclusion criteria.

Step 2b: Calculate the facility observed score by dividing the results of step 2a by the results of step 1.

Step 3: Calculate the national observed score by averaging the scores derived in step 2b across all facilities. 

Step 4: Calculate the expected resident score for each resident (steps 4a and 4b below)

Step 4a: Assign values for the covariate, independence or modified independence in daily decision making on the prior assessment. For the residents included in the denominator for the covariate, assign a ‘0’ if resident interview indicates that the resident is cognitively impaired, or if staff assessment indicates that the resident is moderately to severely impaired in making decisions; and assign a ‘1’ if the resident had no impairment in cognitive status or cognitive skills for daily decision making. Run the logistic regression model.

Specifically, the covariate is calculated as follows:

Covariate = 1 (yes), if the resident had no impairment in cognitive status or cognitive skills for daily decision making as described in 1 and 2 below:

1) Item C1000: Cognitive Skills for Daily Decision Making indicates that the resident made decisions regarding tasks of daily life that were categorized as independent—decisions consistent/reasonable (C1000 = [0]), or modified independence—some difficulty in new situations only (C1000 = [1])

OR

2) Summary Score for Brief Interview for Mental Status indicates intact or borderline cognitive status ([13] ≤ C0500 ≤ [15])

Covariate = 0 (no), if resident interview indicates that the resident is cognitively impaired, or staff assessment indicates that the resident is moderately to severely impaired in making decisions or the values for C1000 and C0500 are missing, as described in 1, 2 and 3 below:

1) Item C1000: Cognitive Skills for Daily Decision Making indicates that the resident made decisions regarding tasks of daily life that were categorized as: moderately impaired—decisions poor; cues/supervision required (C1000 = [2]), or severely impaired—never/rarely made decisions (C1000 = [3])

OR

2) Summary Score for Brief Interview for Mental Status indicates moderate or severe impairment of cognitive status (C0500 ≤ 12)

OR

3) If there is missing data for the Summary Score for Brief Interview for Mental Status (C0500 = [99, -, ^]) and Cognitive Skills for Daily Decision Making (C1000 = [-, ^]).

Covariate = missing if the following is true:

1) No prior assessment is available.

The logistic regression equation is of the form:

[Equation 1] QM triggered (yes=1, no=0) = B0 + B1\*COV

Where:

B0 is the logistic regression constant (B0 = -3.252334),

B1 is the logistic regression coefficient for the covariate (B1 = 1.218125), and

COV is the resident-level score for the covariate (0 or 1)

Step 4b: Calculate the expected resident score for each resident with the following formula:

[Equation 2] Resident-level expected QM score = 1/ [1+e^-x]

Where e is the base of natural logarithms and x is a linear combination of the constant and the logistic regression coefficient times the covariate scores (from Formula [1]). A covariate score will be 1 if the covariate is triggered for that resident, and 0 if the covariate is not triggered.

Step 5: Calculate the facility-level expected QM score by averaging all resident-level expected scores derived in step 4b.  
  
Step 6. Calculate the facility-level adjusted score based on the:

* facility-level observed QM score (step 2b),
* facility-level expected QM score (step 5), and
* national average observed QM score (step 3).

The calculation of the adjusted score uses the following equation:

[Equation 3] Adj = 1/ [1 + e^ -y]

where

Adj is the facility-level adjusted QM score, and

y = (Ln(Obs/(1–Obs)) - Ln(Exp/(1–Exp)) + Ln(Nat/(1–Nat)))

Obs is the facility-level observed QM rate,

Exp is the facility-level expected QM rate,

Nat is the national observed QM rate (Nat= 0.0679302), and

Ln indicates a natural logarithm.

e is the base of natural logarithms

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: LJC03\LJC03\_request\_677.log)

Reference: RTI International MDS 3.0 Quality Measures User’s Manual v11.0. Waltham, MA: RTI International, 2017. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/NursingHomeQualityInits/Downloads/MDS-30-QM-Users-Manual-V11-Final.pdf

**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**.

This is not applicable. This 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 quality measure reports the risk adjusted percentage of residents self-reporting moderate to severe pain among long-stay nursing facility residents. The goal of risk adjustment is to control for differences across facilities in patient characteristics 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 the percentage of patients who are categorized as cognitively independent. The literature, as well as data analysis conducted on this measure, suggests differences in self-reported pain symptoms and pain management between cognitively intact residents and those with a cognitive impairment. While cognitive impairment was found to be negatively associated with self-reported pain (Lukas, Mayer, Fiavola, Topinkova, Gindin, Onder, Bernabei, Nikolaus, & Denkinger, 2013) and receipt of appropriate pain management (Achterberg, Pieper, van Dalen-Kok, WM de Waal, Husebo, Lautenbacher, Kunz, Scherder, & Corbett, 2013; Corbet, Husebo, Malcangio, Staniland, Cohen-Mansfield, Aarsland, & Ballard, 2012; Zwackhalen, van’t Hof, & Hamers, 2012), there is evidence to suggest this is partially due to the residents’ inability to report pain, (Achterberg, et al., 2013; Herr, 2011) rather than due to a decline in pain. Thus, this measure is risk-adjusted to allow percentage of residents self-reporting pain to be compared across nursing facilities, which may have differing proportions of residents with cognitive independence.

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

**Clinical Factors**

Past RTI analysis (previously submitted to NQF) identified potential clinical risk adjustors, selected based on literature review and clinical relevance, including oldest old (age ≥ 85 years), make self understood, stroke, arthritis, and asthma/COPD. Bivariate analyses were conducted to examine these identified potential risk factors. Clinical review of results from the bivariate analyses was then conducted to determine if there was a clinical rationale to risk adjust for these factors. Two criteria were then applied in the selection of risk factors: 1) at least 5% of residents have the described characteristic; and 2) bivariate analyses showed that that the examined resident characteristics were significantly associated with the measured outcome. Selected characteristics were then systematically included in a series of multivariate logistic regression models to test their independent importance in predicting the pain outcome and in improving the predictive power of the model. We made decisions to retain or drop risk adjustors based on sample size, regression coefficient, significance level, and clinical relevance, and predictive power of the model.

The first model tested included ‘independence’ or ‘modified independence’ in daily decision making on the prior assessment as the predictor variable and numerator inclusion in this as measure as the predicted variable. The odds ratio for cognitive independence is 2.91 (*p* < .001), and the c-statistic is 0.629 (pseudo *R*2 = 0.040). Adding oldest old and make self understood into the model moderately increased the c-statistic to 0.665 (pseudo *R*2 = 0.050). Among the three covariates, independence or modified independence in daily decision making on the prior assessment was the strongest predictor, suggested by an adjusted odds ratio (OR) of 2.3 (*p* < .001). The adjusted ORs for oldest old and make self understood were 0.7 (*p* < .001) and 1.7 (*p* < .001), respectively. Introducing disease diagnoses (i.e., stroke, arthritis, and asthma/COPD) as covariates to the model slightly increased the c-statistic to 0.676 (pseudo *R*2 = 0.050). Despite being statistically significant, the presence of stroke influenced the risk of pain to a very small extent, indicated by an adjusted OR of 0.95 (*p* < .001). Arthritis and asthma/COPD both had a moderate impact on risk of pain (adjusted OR = 1.6 and 1.4, respectively; *p* < .001). In summary, the risk adjustment currently used for this QM showed acceptable predictive power for resident episodes in Quarter 3, 2011. The current covariate independence or modified independence in daily decision making on the prior assessment appeared to be the strongest predictor among other potential candidate risk adjustors, which only moderately improved the risk adjustment model’s predictive power.

Results of these analyses supported the inclusion of the covariate, independence or modified independence in daily decision making, as individuals with no impairments appear to systematically report pain more frequently than those who do have cognitive impairments. This testing form contains updated information on the currently specified risk adjustment model for this measure, which risk adjusts for cognitive independence only. For the current specification, we also provide the results of regression diagnostics assessing model performance, including model fit and calibration statistics. In addition, we note that, by including cognitive independence in the risk adjustment model, we restrict the measure to only those residents with both a target and a prior assessment, which reduces the number of residents included in the denominator of this measure by less than 1%, attenuating concerns regarding the reportability of this measure. Furthermore, based on logistic regression model results and model fit and calibration statistics, we conclude that the risk adjustment model that includes cognitive independence only is parsimonious, and additional risk adjustment for social risk factors offers very little improvement in predictive ability of the model.

RTI analysis of Q3, 2011 MDS 3.0 Data (programming reference: quarter\_3\_4\_complete\ntz12\_request\_011.log)

RTI analysis of Q2, 2017 MDS 3.0 Data (programming reference: LJC04/LJC04\_request\_677.log)

**Social Risk Factors**

In addition, since this measure was last endorsed, NQF has lifted the previous restriction on risk-adjusting for social risk factors on a trial basis. Therefore, we conducted analyses on the feasibility of including several social risk factors in the previously specified risk-adjustment model included in this testing form. We identified potential social risk factors that may be related to self-reported pain based on a review of the current literature.

We conducted a series of analyses to assess whether these patient-level social risk factors would be appropriate risk adjustors in multivariate models. We examined the distribution of each social risk factor across facilities. If a social risk factor was evenly distributed across facilities (i.e., it did not tend to cluster in some facilities), it would not be necessary to risk adjust for this social risk factor. We also conducted bivariate analyses of this quality measure for the social risk factors identified in the literature to assess whether each social risk factor appeared to have a relationship with self-reported pain. These selected characteristics were then included in separate multivariate logistic regression models to examine the impact of adjusting for each social risk factor, in addition to cognitive independence.

The results of analyses to assess the appropriateness and feasibility of risk-adjusting for these social risk factors are described below in 2b3.3a. Based on the results of our analyses, which are described below, we do not recommend risk adjusting this measure for social risk factors.

***Risk Adjustor Included in the Final Model***

The risk adjustor selected for inclusion in the final model, cognitive independence, is described in detail in 2b3.1.1.

Achterberg, W., Pieper, M., van Dalen-Kok, A., WM de Waal, M., Husebo, B., Lautenbacher, S., Kunz, M., Scherder, E., & Corbett, A. (2013) Pain management in patients with dementia. *Clinical Interventions in Aging, 8*, 1471-1482.

Corbett, A., Husebo, B., Malcangio, M., Staniland, A., Cohen-Mansfield, J., Aarsland, D., Ballard, C. (2013). Assessment and treatment of pain in people with dementia. Nature Reviews Neurology, 8, 264-274.

Lukas, A., Mayer, B., Fiavola, D., Topinkova, E., Gindin, J., Onder, G., Bernabei, R., Nikolaus, T., & Denkinger, M. (2013) Pain characteristics and pain control in European nursing homes: Cross-sectional and longitudinal results from the services and health for elderly in long Term care (SHELTER) study. *Journal of the American Medical Directors Association, 14*(6), 421-428.

Herr, K. (2011) Pain assessment strategies in older patients. *The Journal of Pain, 12*(3), S3-S13.

Zwakhalen, S., van’t Hof, C., & Hamers, J. (2012) Systematic pain assessment using an observational scale in nursing home residents with dementia: Exploring feasibility and applied interventions. *Journal of Clinical Nursing, 21*(21), 3009-3017.

**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)**

Based on a review of the current literature, we identified several social risk factors that may be related to self-reported pain and could be measured by items available in the MDS 3.0. While the literature reported mixed results on the relationship between socio-economic status (SES) and self-reported pain, as well as between marital status and self-reported pain, overall, evidence suggests that self-reported pain is more likely to be associated with lower socio-economic status (Kuntz, et al., 2017; Johannes, Le, Zhou, Johnston, & Dworkin, 2010) and not being married [i.e., widowed, divorced, separated, and never married (Johannes, et al., 2010; Sjogren, Ekholm, Peuckmann, & Gronbaek, 2009)]. The evidence suggests that this may be due to low levels of self-involvement in healthcare and living in a less affluent area with access to fewer supports for those with lower SES (Brekke, Hjortdahl, & Kvien, 2002). Marriage has been shown to be associated with better health and longer life expectancy due to the social and societal connections from which married individuals benefit, possibly explaining the lower rates of self-reported pain among married individuals than among non-married individuals (Wade, Hart, Wade, Bajaj, & Price, 2013). However, one study to note by Sawyer and colleagues (2007) found that self-reported pain was higher among married nursing home residents than non-married residents. Due to the mixed results in the literature, SES and marital status were examined. Given the available MDS data, we examined the effect of eligibility for Medicaid as an indicator of socio-economic status.

Likewise, the literature points to lower prevalence of self-reported pain among the oldest nursing home residents (Clement, Bradley, & Lin, 2009; Takai, Yamamoto-Mitani, Okamoto, Koyama, & Honda, 2010) and among male residents (Lukas, Mayer, Fiavola, Topinkova, Gindin, Onder, Bernabei, Nikolaus, & Denkinger, 2013; Sawyer, et al., 2007; Takai, et al., 2010); thus, we also explored the impact of being 85 years or older (‘oldest old’) and gender on this measure. The literature attributes much of the differences in self-reported pain between groups to psychosocial factors. The decline in self-reported pain among the oldest old may be related to cohort differences or an acceptance of pain as part of the aging process (Zarit, Griffiths, & Berg, 2004). Differences in self-reported pain between men and women are thought to be associated with socially learned reactions (Bartley & Fillingim, 2013; Robinson, Riley, Myers, Papa, Wise, Waxenberg, & Fillingim, 2001), such as pain-related expectations, sex role beliefs, and pain coping strategies (Fillingim, 2000).

We created binary variables for each social risk factor described above as follows:

* Medicaid eligibility: defined from Item A0700 (Medicaid Number) in the MDS. Medicaid eligibility is defined as 1 if the resident has a Medicaid number or if a Medicaid number is pending, 0 if Medicaid number = “N”, and missing if Medicaid number is missing.
* Oldest old: defined from Item A0900 (Birth Date) in the MDS. Oldest old is defined as 1 if the resident is age 85 or older and 0 if otherwise. Birth Date is not missing on any assessment in the sample.
* Not married: defined from Item A1200 (Marital Status) in the MDS. Not married is defined as 1 if the resident is never married, widowed, separated, or divorced; 0 if the resident is married; and missing if marital status is missing on the target assessment.
* Gender: defined from item A0800 (Gender) in the MDS. Male is defined as 1 and Female as 0. Gender is not missing on any assessment in the sample.

We would also like to note that while there are observed differences in self-reported pain by race/ethnicity, we did not consider race as a potential social risk factor per NQF guidance on the use of race as an indicator of socioeconomic status. Detailed analyses on the impact of race on the percentage of residents who self-report moderate to severe pain can be found in the Measure Submission Form.

Bartley, E. & Fillingim, R. (2013) Sex differences in pain: a brief review of clinical and experimental findings. British Journal of Anesthesia, 111(1), 52-58.

Brekke, M., Hjortdahl, P., & Kvien, T. (2002) Severity of musculoskeletal pain: relations to socioeconomic inequality. Social Science & Medicine, 54(2), 221-228.

Clement, J., Bradley, C., & Lin, C. (2009) Organizational characteristics and cancer care for nursing home residents. *Health Services Research, 44*(6), 1983-2003.

Fillingim, R. (2000) Sex, gender, and pain: Women and men really are different. Current Review of Pain, 4(1), 24-30.

Johannes, C., Le, K., Zhou, X., Johnston, J., & Dworkin, R. (2010). The prevalence of chronic pain in united states adults: Result of an internet-based survey. *The Journal of Pain, 11*(11), 1230-1239.

Kunz, B., Hoebel, J., Fuchs, J., Neuhauser, H., Lampert, T. (2017). Social inequalities in the prevalence of chronic back pain among adults in Germany. *Bundesgesundheitsblatt, 60*(7), 783-791.

Lukas, A., Mayer, B., Fiavola, D., Topnkova, E., Gindin, J., Onder, G., Bernabei, R., Nikolaus, T., & Denkinger, M. (2013) Pain characteristics and pain control in European nursing homes: Cross-sectional and longitudinal results from the services and health for elderly in long term care (SHELTER) study. *Journal of the American Medical Directors Association, 14*(6), 421-428.

Robinson, M., Rile, J., Myers, C., Papa, R., R., Wise, E., Waxenberg, L., & Fillingim, R. (2001) Gender role expectations of pain: Relationship to sex differences in pain. The Journal of Pain, 2(5), 251-257.

Sawyer, P., Lillis, J. P., Bodner, E., & Allman, R. (2007) Substantial daily pain among nursing home residents. *Journal of the American Medical Directors Association, 8*(3), 158-165.

Sjogren, P., Ekholm, O., Peuckmann, V., & Gronbaek, M. (2009) Epidemiology of chronic pain in Denmark: An update. European Journal of Pain, 13(3), 287-292.

Takai, Y., Yamamoto-Mitani, N., Okamoto, Y., Koyama, K., Honda, A. (2010). Literature review of pain prevalence among older residents of nursing homes. *Pain Management Nursing, 1*(4), 209-223.

Wade, J. B., Hart, R., Wade, J. H., Bajaj, J., & Price, D. (2013) The relationship between marital status and psychological resilience in chronic pain. Pain Research and Treatment, 2013, 928473.

Zarit, S., Griffiths, P., & Berg, S. (2004) Pain perceptions of the oldest old: A longitudinal study. The Gerontologist, 44(4), 459-468.

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**The current risk adjustment model includes one covariate, cognitive independence, which is a significant predictor of self-reported pain. For the covariate risk adjustment (indirect standardization), the measure is adjusted for independence or modified independence in daily decision making (C1000=0 or 1) for residents evaluated with the staff assessment, or no cognitive impairments (C0500>12) coded on the prior MDS assessment (referred to hereafter as cognitive independence).

First, using data from Quarter 2, 2017, we examined the percentage of long-stay residents who were cognitively independent who self-reported moderate to severe pain, compared to those who were cognitively impaired, and used Chi-Squared tests to determine whether these differences were statistically significant, as shown in ***Table 3***. There is a large, significant difference in the proportion of residents who self-report moderate to severe pain by cognitive independence: among residents who are cognitively independent, 11.6% self-report moderate to severe pain, and among residents who are cognitively impaired, 3.7% self-report moderate to severe pain (χ2(1) >19,000, p <.001).

**Table 3. Percent of Residents Who Self-Report Moderate to Severe Pain by Cognitive Independence**

| Resident characteristic (% of all long stay residents) | Frequency of residents who self-report moderate to severe pain (n) | Percentage of residents who self-report moderate to severe pain, % | Pearson chi2 P-value |
| --- | --- | --- | --- |
| Cognitively independent (39.13%) | 37,652 | 11.6% | <0.001 |
| Cognitively impaired (60.87%) | 18,857 | 3.7% |  |

Source: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC03/LJC03\_request\_677.log)

Next, we examined the facility-level distribution of the percentage of residents who are cognitively independent (presented in ***Table 4***). The facility-level mean was 39.3%, with a standard deviation of 13.9%.

**Table 4. Distribution of Percentage of Residents who are Cognitively Independent Across facilities**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Resident characteristics | Facilities (*k)* | Mean % of residents | Std dev. | 10th  percentile | 25th  percentile | 50th  percentile | 75th  percentile | 90th  percentile |
| Cognitively independent | 13,691 | 39.3% | 13.9% | 22.0% | 30.3% | 39.1% | 48.1% | 56.8% |

Source: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC03/LJC03\_request\_677.log)

The odds ratio for cognitive independence is 3.38 and is statistically significant at the 0.05 level [95% CI = [3.32, 3.44]. The odds of self-reporting moderate to severe pain are 2.38 times higher among those with no impairment in cognitive status, compared to those with cognitive impairments.

The facility-level mean unadjusted percentage of long-stay residents self-reporting moderate to severe pain was 7.5% while the risk adjusted facility-level mean was 7.4% (prior to applying public reporting restrictions for facility size).Although the overall facility-level means for the risk-adjusted and unadjusted quality measures are not substantially different, the literature review and statistical evidence (including the relatively large and statistically significant odds ratio for the covariate, as well as the model fit statistics described in 2b3.6) indicate that this risk adjustment model is valid.

Results of the statistical analyses to examine social risk factors as potential risk adjustors are detailed in 2b3.4b.

**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.**

First, we examined the percentage of long-stay residents with each social risk factor who self-reported moderate to severe pain compared to those without that social risk factor, and used Chi-Squared tests to determine whether these differences were statistically significant, as shown in ***Table 5***.

While all differences were statistically significant, the differences across subpopulations are mostly small. Among residents who are eligible for Medicaid, 6.9% self-report moderate to severe pain and, among those ineligible for Medicaid, 6.3% report pain ((*χ*2(1) = 46.6, *p* < .001); among unmarried residents, 6.9% reported moderate to severe pain, whereas 6.5% of married residents reported pain (*χ*2(1) = 26.7, *p* < .001). The age and gender related differences we found were relatively larger. For residents aged 85 years or older, only 4.6% reported moderate to severe pain, compared with 8.3% of younger residents (*χ*2(1) = 4,600, *p* < .001). Whereas 6.1% of the male residents reported moderate to severe pain, 7.2% of the female residents reported this level of pain (*χ*2(1) = 345.8, *p* < .001).

Differences in percentages of individuals self-reporting moderate to severe pain were in the expected direction: younger residents, non-married residents, female residents, and Medicaid-eligible residents have higher proportions of individuals who self-reported moderate to severe pain.

**Table 5. Percent of Residents Who Self-Report Moderate to Severe Pain by Social Risk Factors**

| Resident characteristic (% of all long stay residents) | Frequency of residents who self-report moderate to severe pain (n) | Percentage of residents who self-report moderate to severe pain (%) | Pearson chi2 P-value |
| --- | --- | --- | --- |
| Age |  |  |  |
| ≥ 85 (40.71%) | 15,394 | 4.6% | <0.001 |
| < 85 (59.29%) | 41,115 | 8.3% |  |
| Sex |  |  |  |
| Male (35.11%) | 17,805 | 6.1% | <0.001 |
| Female (64.89%) | 38,704 | 7.2% |  |
| Marital Status |  |  |  |
| Not married (82.43%) | 46,053 | 6.9% | <0.001 |
| Married (17.57%) | 9,272 | 6.5% |  |
| Medicaid |  |  |  |
| Medicaid (87.49%) | 46,078 | 6.9% | <0.001 |
| Non-Medicaid (12.51%) | 6,021 | 6.3% |  |

Source: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC03/LJC03\_request\_677.log)

Next, we examined the facility-level distribution of each social risk factor (presented in ***Table 6***). Among the tested social risk factors, there was considerable cross-facility variation; this variation was relatively smaller in the percentage of residents who are not married.

In addition, we used the facility-level distributions to assess whether each social risk factor may be appropriate to include in risk adjustment models. It may not be appropriate to risk adjust for social risk factors that:

1. Have either a very high or very low prevalence
2. Are evenly distributed across facilities (i.e., they did not cluster in some facilities)

Overall, the facility-level mean of Medicaid-eligible residents is high (87.0%), and the distribution is right-skewed; however, this high proportion of Medicaid-eligible residents may be related to the target population for this quality measure (i.e., long-stay nursing facility residents), and not necessarily indicative of socio-economic status. The facility-level mean of residents who are not married is also high (82.0%); this may also be related to the target population, as those who are married may have more social supports and may be able to remain in the community.

**Table 6. Distribution of Percentage of Residents with Select Resident Characteristics Across Facilities**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Resident characteristics | Facilities (*k)* | Mean % of residents | Std dev. | 10th  percentile | 25th  percentile | 50th  percentile | 75th  percentile | 90th  percentile |
| Age ≥ 85 | 13,691 | 41.9% | 20.1% | 15.4% | 26.9% | 41.7% | 56.3% | 68.9% |
| Male | 13,691 | 34.1% | 13.4% | 18.9% | 25.0% | 32.3% | 41.3% | 51.4% |
| Not married | 13,689 | 82.0% | 8.0% | 71.8% | 77.3% | 82.6% | 87.5% | 91.7% |
| Medicaid eligible | 13,662 | 87.0% | 19.3% | 62.7% | 82.1% | 94.9% | 100.0% | 100.0% |

Source: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC03/LJC03\_request\_677.log)

***Tables 5*** and ***6*** above indicate that there is a high prevalence of residents who are eligible for Medicaid and who are not married in the target population, that the relationship between these resident characteristics and self-reported pain is in the hypothesized direction, and that the facility-level means of these resident characteristics are also high. Because these variables may be associated with the target population (i.e., long stay nursing facility residents) and may be less indicative of socio-economic status in this population, these social risk factors were not further considered as potential risk adjustors for this measure.

Being male was associated with self-reported pain in the expected direction, with the proportion of residents who trigger the numerator for this measure being significantly higher among females than males; thus, we further considered male gender as a potential risk adjustor for this measure. Age of 85 years and older was also associated with self-reported pain in the hypothesized direction; the difference in the proportion of residents who trigger the numerator by age is large and statistically significant and, therefore, we further considered age of 85 years and older as a potential risk adjustor for this measure.

Based on the results of bivariate analyses, we examined three additional logistic regression models:

1. Risk adjusted for cognitive independence and oldest old (Model 1)
2. Risk adjusted for cognitive independence and male gender (Model 2)
3. Risk adjusted for cognitive independence, oldest old, and male gender (Model 3)

In Model 1, the odds ratio for cognitive independence is 3.12 [95% CI = [3.06, 3.18], and the odds ratio for oldest old is 0.67 [95% CI = [0.66, 0.68]. Both odds ratios are statistically significant at the 0.05 level. Similar to the results of the current risk adjustment model, the odds of self-reporting moderate to severe pain are over 2.12 times higher among those with no impairment in cognitive status, compared to those with cognitive impairments. Consistent with previous studies, the odds of self-reporting moderate to severe pain are 33% lower among residents aged 85 and older, compared to younger residents.

In Model 2, the odds ratio for cognitive independence is 3.43 [95% CI = [3.37, 3.49], and the odds ratio for male gender is 0.78 [95% CI = [0.76, 0.79]. Both odds ratios are statistically significant at the 0.05level. Similar to the results of the current risk adjustment model, the odds of self-reporting moderate to severe pain are over 2.43 times higher among those with no impairment in cognitive status, compared to those with cognitive impairments. In addition, the odds of self-reporting moderate to severe pain are 22% lower among male residents than female residents.

In Model 3, the odds ratio for cognitive independence is 3.13 [95% CI = [3.08, 3.19], the odds ratio for oldest old is 0.62 [95% CI = [0.61, 0.63], and the odds ratio for male gender is 0.71 [95% CI = [0.69, 0.72]. All odds ratios are statistically significant at the 0.05 level. Similar to the results of the current risk adjustment model, the odds of self-reporting moderate to severe pain are over 2.13 times higher among those with no impairment in cognitive status, compared to those with cognitive impairments. As in Model 1, the odds of self-reporting moderate to severe pain are 38% lower among residents aged 85 and older, compared to younger residents. As in Model 2, the odds of self-reporting moderate to severe pain are 29% lower among male residents than female residents.

We then compared the model fit and calibration for these two models to those for the current risk adjustment specification. When we additionally risk adjust for oldest old, the c-statistic increases slightly, by 0.02 (from 0.65 to 0.67), compared to the c-statistic for the current risk adjustment model. Similarly, when we additionally risk adjust for gender, the c-statistic increases slightly, by 0.01 (from 0.65 to 0.66). In addition, when we risk adjust for cognitive independence, oldest old, and gender, the c-statistic increases by 0.03 (from 0.65 to 0.68), compared to the c-statistic for the current risk adjustment model.

***Table 7*** provides further comparison of the current risk adjustment model and alternate model specifications.

Table 7. Comparison of Model Performance for Alternate Risk Adjustment Specifications, NQF #0677

| Model Covariates | Hosmer-Lemeshow Chi2, *P*-value | AIC | BIC | Pseudo *R*2 | Log Likelihood |
| --- | --- | --- | --- | --- | --- |
| Cognitive independence | -- | 394,307.5 | 394,330.8 | 0.0453 | -197,151.8 |
| Model 1 | 2.03, 0.362 | 392,654.3 | 392,689.2 | 0.0493 | -196,324.1 |
| Model 2 | 80.94, <0.001 | 393,583.4 | 393,618.3 | 0.0471 | -196,788.7 |
| Model 3 | 15.40, 0.004 | 391,327.4 | 391,373.9 | 0.0526 | -195,659.7 |

NOTES: “-- “indicates that the Hosmer-Lemeshow test has two quantiles, and the significance of the test statistic cannot be computed.

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

Model 1: Risk adjusted for cognitive independence and oldest old

Model 2: Risk adjusted for cognitive independence and male gender

Model 3: Risk adjusted for cognitive independence, oldest old, and male gender

Source: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC04/LJC04\_request\_677.log)

Based on the c-statistic and these model calibration statistics, overall, Model 3 appears to have the best fit and calibration. However, we also note that the improvements to model fit and calibration by including these additional social risk factors are very slight, and there is almost no practical improvement in adding these risk adjustors to the risk adjustment model.

In addition, we examined the correlation between quality measure scores, comparing the quality measure scores under the risk adjustment model specification for Models 1-3 to the current risk adjustment model (adjusted for cognitive independence only). The quality measure scores calculated following the risk adjustment specifications in Models 1-3 are very highly correlated with the quality measure score under the current risk adjustment model (>0.996 for all three models risk-adjusted for social risk factors, *p* < .001). The high degree of correlation between the risk-adjusted QM scores suggests that, for most nursing facilities, the current risk adjustment model and alternate model specifications are not significantly different.

Source: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC04/LJC04\_request\_677.log)

**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 self-reported pain among long-stay nursing home residents. After running each of the logistic regression models, we conducted regression diagnostics to assess model performance, examining predictive ability, as well as model fit and calibration.

We used the c-statistic to examine the discrimination of the statistical risk model, and we used the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), pseudo *R*2, and log-likelihood to examine the statistical risk model calibration. We also used the Hosmer-Lemeshow test for goodness of fit; however, with only 1 predictor in the risk adjustment model, the Hosmer-Lemeshow test only has two quantiles, and the significance of the test statistic cannot be computed. In addition, we present the receiving operating characteristic (ROC) curve in 2b3.8.

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

The risk adjustment model currently used for this quality measure had acceptable predictive power on the data in Quarter 2, 2017. The multivariate logistic regression model with independence or modified independence in daily decision making on the prior assessment as the only covariate yielded a c-statistic of 0.65 and a pseudo *R*2 of 0.045, values typical for risk adjustment models of nursing facility outcomes based on MDS assessment data (Mukamel, 1997).

SOURCE: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC04/LJC04\_request\_677.log)

Mukamel, D. B. (1997). Risk adjusted outcome measures and quality of care in nursing homes. Medical Care, 35, 367–385

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

The risk adjustment model currently used for this quality measure performed similarly to models with additional risk adjustors when using data from Quarter 2, 2017. ***Table 8*** below presents the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood for the current risk adjustment model. These statistics are also presented for all risk adjustment models in ***Table 7*** above. In general, when comparing models, smaller values for AIC or BIC indicate better model fit; thus, models with smaller values for the information criterion are considered preferable. The information criterion in ***Tables 7*** and ***8*** can be used to compare alternate model specifications.

Table 8. Statistical Risk Model Calibration Statistics

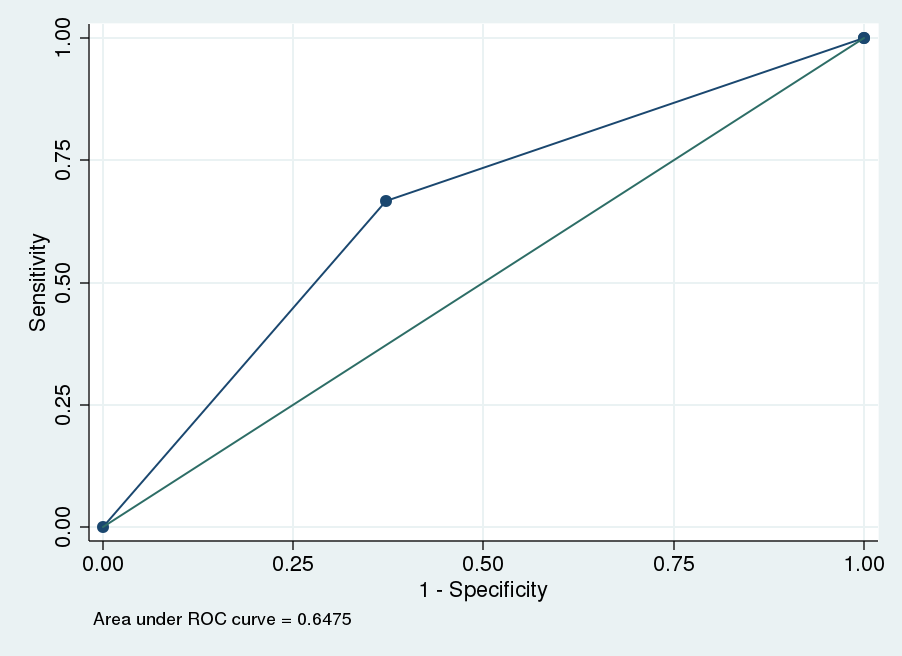
| Model Covariates (N) | AIC | BIC | Log Likelihood |
| --- | --- | --- | --- |
| Cognitive independence (831,869) | 394,307.5 | 394,330.8 | -197,151.8 |

SOURCE: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC04/LJC04\_request\_677.log)

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

The receiver operating characteristic (ROC) curve is presented in ***Figure 4*** below. With one binary predictor (i.e., one possible cut-point for this test), the ROC curve is not smooth. Using the current model specification, the area under the curve indicates acceptable model performance, with alternate specifications that include additional risk adjustors performing slightly better.

**Figure 4. ROC Curve, QM #0677 Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay), Quarter 2, 2017**



Source: RTI analysis of MDS 3.0 Data, Q2 2017 (programming reference: LJC04/LJC04\_request\_677.log)

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

Not applicable. This measure is not stratified.

**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 that the current risk adjustment model has reasonable predictive power; additional risk adjustment for social risk factors offers little practical improvement to the risk adjustment model.

**2b3.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

RTI described the current variability in the facility-level quality measure scores.

RTI also examined proportions of facilities with scores for this measure that are significantly different from national facility-level mean, stratified by facility denominator size. For this analysis, statistical significance was determined using 95% confidence intervals: a facility’s quality measure score was significantly different from the national mean if the national mean was not included in the facility’s 95% confidence interval. High-performing facilities should have scores that are significantly below average, and scores of low-performing facilities should be significantly above average. We stratified the analysis by facility denominator size to examine whether this feature of the measure varies by size.

**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 9*** describes the current variability in the quality measure scores of facilities nationally. We find that the mean facility-level score for this quality measure was 7.2% in Quarter 2, 2017 with a median score of 4.9%, indicating a slight positive skew. The interquartile range for this measure was 9.6%. Among facilities who were eligible to publicly report this measure, 22.7% had perfect scores of 0%; as shown in ***Figure 3*** in Section 2b1, the national mean facility-level score on this measure has decreased over time, suggesting that facilities may have improved in pain management after this measure was publicly reported.

Table 9  
National Facility-Level Score Distribution, NQF #0677 Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *k* | Mean score | Std dev. | 10th  percentile | 25th  percentile | 50th  percentile | 75th  percentile | 90th  percentile | % of  facilities with  “perfect scores” | Interquartile  range |
| 13,691 | 7.2% | 7.6% | 0.0% | 1.2% | 4.9% | 10.8% | 17.8% | 22.7% | 9.6% |

NOTES: *k* = number of facilities that meet minimum requirements for public reporting this quality measure.

Analysis date: 4/5/2018

SOURCE: RTI analysis of MDS 3.0 data for Quarter 2, 2017 (programming reference: rn01/ ac356\_request\_q2627\_677.log; rn-1/ac362\_request\_q2627\_677.log)

***Table 10*** shows the proportions of facilities that score statistically significantly higher or lower than the national facility-level mean in Quarter 2, 2017. For this analysis, statistical significance was determined using 95% confidence intervals: a facility’s quality measure score was significantly different from the national mean if the national mean was not within the facility’s 95% confidence interval.

Overall, almost half (46.2%) of facilities were significantly different from the national mean in Quarter 2, 2017, indicating that there are meaningful differences in facility-level scores for this measure. We also stratified the data by the facility denominator size to allow us to examine the relationship between facility size and the reliability of facility scores. The proportions of facilities with scores that are significantly different from the national mean vary as a function of the number of residents included in the denominator for this measure; in general, the percentage of facilities which have scores that are statistically significantly different from the mean increase with the number of residents. Increases in the facility-level sample size lead to reductions in the standard error of facility-level scores, thus, it appears that changes in the reliability of this measure for larger facilities are due to the greater statistical reliability that accompanies increased sample size.

Table 10  
Proportion of Facilities with Scores Significantly Different from the National Facility-Level Mean, Stratified by Facility Denominator Size for NQF #0677 Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)

| Decile of denominator size in residents | *k* | | | Number of facilities with 95% confidence interval lower than national mean (%) | | Number of facilities with 95% confidence interval higher than national mean (%) | | Total number of facilities with scores significantly different from mean (%) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1st Decile (*n*= 20 to 27) | | 1,419 | 441 (31.1%) | | 100 (7.0%) | | 541 (38.1%) | |
| 2nd Decile (*n* = 28 to 33) | | 1,352 | 378 (28.0%) | | 107 (7.9%) | | 485 (35.9%) | |
| 3rd Decile (*n* = 34 to 39) | | 1,383 | 381 (27.5%) | | 123 (8.9%) | | 504 (36.4%) | |
| 4th Decile (*n* = 40 to 46) | | 1,518 | 504 (33.2%) | | 156 (10.3%) | | 660 (43.5%) | |
| 5th Decile (*n* = 47 to 52) | | 1,289 | 446 (34.6%) | | 125 (9.7%) | | 571 (44.3%) | |
| 6th Decile (*n* = 53 to 59) | | 1,380 | 489 (35.4%) | | 144 (10.4%) | | 633 (45.9%) | |
| 7th Decile (*n* = 60 to 67) | | 1,380 | 489 (35.4%) | | 159 (11.5%) | | 648 (47.0%) | |
| 8th Decile (*n* = 68 to 77) | | 1,248 | 513 (41.1%) | | 117 (9.4%) | | 630 (50.5%) | |
| 9th Decile (*n* = 78 to 97) | | 1,367 | 631 (46.2%) | | 144 (10.5%) | | 775 (56.7%) | |
| 10th Decile (*n* = 98 to 552) | | 1,355 | 725 (53.5%) | | 155 (11.4%) | | 880 (64.9%) | |
| Total | | 13,691 | 4,997 (36.5%) | | 1,330 (9.7%) | | 6,327 (46.2%) | |

NOTE: *k* = number of facilities that meet minimum requirements for public reporting this quality measure.

Analysis date: 4/5/2018

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: rn01\ac364\_request\_q2627\_677.log)

**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 analyses show that the quality measure score varies enough to make meaningful distinctions between high- and low-quality facilities. Moreover, the quality measure scores vary enough from the national mean that there are meaningful differences in facility-level scores for this measure.

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

**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*)

RTI analyzed the effects of missing data on this measure in the following ways:

1. We report summary statistics for the facility-level distribution of missing data rates for items used in the calculation of the long-stay pain measure, both overall and stratified by quality measure score quartile.
2. We analyzed whether missing data on pain items varied systematically by several resident-level characteristics which are associated with self-reported pain.
   1. Specifically, we examined if missing data on pain items J0300 (Pain Presence), J0400 (Pain Frequency), or J0600 (Pain Intensity) varied systematically on the following characteristics which our analyses show are related to self-reported pain: age greater than or equal to 85, gender, and a BIMS score of less than or equal to 12 (indicating cognitive impairment).

**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*)

1. Among all long-stay residents in Quarter 2, 2017 (*n*=1,120,814), 8.6% (*n*=96,793) had missing data for at least one of the items necessary to calculate the pain measure. Similarly, the mean facility-level percentage of residents with missing data for any of the items necessary to calculate the pain measure was 8.2% (***Table 11***).

***Table 11*** shows facility-level summary statistics for the distribution of missing data rates for items used to calculate this measure, both overall and stratified by quartile of quality measure scores.

Table 11. Distribution of Facility-Level Missing Rate by Measure Score Quartile, NQF #0677 Percent of Residents Who Self-Report Moderate to Severe Pain (Long Stay)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Quality Measure Score Quartile** | **Facilities (k)** | | **Mean** | **Std dev.** | **10th** | **25th** | **50th** | **75th** | **90th** |
| 0–25 (Best)† | | 3,423 | 7.7% | 6.8% | 1.8% | 3.6% | 6.4% | 9.8% | 14.3% |
| 26–50 | | 3,425 | 7.7% | 6.0% | 2.0% | 3.9% | 6.6% | 9.9% | 14.2% |
| 51–75 | | 3,422 | 8.4% | 7.1% | 2.1% | 4.0% | 6.9% | 10.5% | 15.5% |
| 76–100 (Worst)† | | 3,421 | 9.0% | 7.4% | 2.3% | 4.5% | 7.4% | 11.5% | 16.9% |
| Total | | 13,691 | 8.2% | 6.9% | 2.1% | 4.0% | 6.8% | 10.4% | 15.2% |

NOTES: †Facilities with scores in the lowest part of the distribution are considered to have better performance relative to facilities with scores in higher parts of the distribution.

Analysis date: 4/5/2018

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: rn16/rn16\_request\_ls.log; av\_db361/new/av\_db361\_rerun\_new.log)

Our analysis showed that the average missing data rate across facilities was 8.2% (shown in the “Total” row of ***Table 11***) on items used to construct this measure, with a greater percentage of facilities having lower missing data rates and relatively fewer facilities having higher missing data rates. ***Table 11*** indicates that only 10% of facilities had missing data rates of 15.2% or higher.

In addition, RTI examined the relationship between missing data for items used to calculate this measure by quality measure score quartile. ***Table 11*** shows the mean facility-level missing rate for items used to calculate this measure increases with the score quartile, from the quartile containing the lowest (best) scores (missing rate: 7.7%) to the quartile containing the highest (worst) scores (9.0%). This pattern is also shown by the significant but weak correlation between missing data and quality measure scores (*r* = .090, *p* < .001).

This analysis addresses the potential concern that missing data in the pain items would lead to under-reporting residents in pain, resulting in lower (better) scores on this measure. Our analysis demonstrated that this is not an issue for this measure. The positive relationship between missing rate and percentage of residents who self-report pain suggests that facilities with poorer pain management may not do well in assessing and documenting pain.

1. ***Table 12*** summarizes the results of RTI’s analysis of whether missing data varied across selected resident characteristics related to self-reported pain.

Specifically, we analyzed whether missing data on pain items J0300, J0400, or J0600 varied systematically on the following characteristics: age greater than or equal to 85, gender, and a BIMS score of less than or equal to 12 (indicating cognitive impairment), and used Chi-Squared tests to determine whether these differences were statistically significant. ***Table 12*** summarizes the results of this analysis.

Table 12  
 Frequency of Missing Data by Select Resident Characteristics Among Long-Stay Residents

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Any missing data on J0300, J0400, or J0600 | Age | | Gender | | BIMS Score | |
| Age < 85 | Age ≥ 85 | Female | Male | BIMS > 12 | BIMS ≤ 12 |
| Percent of Residents (%) | 11.3% | 8.6% | 9.5% | 11.5% | 2.1% | 3.0% |
| Frequency of Residents (n) | 64,418 | 32,375 | 58,050 | 38,743 | 6,910 | 13,987 |
| Pearson chi2 P-value | <0.001 | | <0.001 | | <0.001 | |

NOTES: J0300 = Pain Presence; J0400 = Pain Frequency; J0600 = Pain Intensity; Please note residents who were excluded because they were unable to participate in the pain assessment interview (i.e., they were rarely or never understood (J0200 = [0]) or they were unable to answer when asked, “Have you had pain or hurting at any time in the last 5 days? (J0300 = [9])) are not considered to have missing data and are not presented in this table; rather we consider them excluded because they were unable to self-report.

Analysis date: 12/5/2017

SOURCE: RTI analysis of Q2 2017 MDS 3.0 data (programming reference: rn16/rn16\_request\_ls.log)

***Table 12*** shows that, while all differences were statistically significant, rates of missing data on pain items are not drastically different among individuals by age, gender and cognitive impairment. For residents aged 85 years or older, 8.6% had missing data on J0300, J0400, or J0600, compared to 11.3% of younger residents (*χ*2(1) = 3,300, *p* < .001). Whereas 11.5% of the male residents had missing data, 9.5% of the female residents had missing data (*χ*2(1) = 4,100, *p* < .001). In addition, 2.1% of residents with a BIMS score greater than 12 had missing data, and 3.0% of residents with a BIMS score ≤12 had missing data (*χ*2(1) = 16,000, *p* < .001).

**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*)

Overall, the mean facility-level missing data rate for this measure was relatively low (8.2%), and only 10% of facilities had missing data rates of 15.2% or higher.

Of the 8.6% of residents excluded from this measure due to missing data on the pain items (*n* = 96,793), 65.0% (*n* = 62,942) had target assessments that were unplanned discharge assessments. If the unplanned discharge item set is used, the resident interview items are not present and there is no opportunity to conduct a resident interview or staff assessment. In these cases, the MDS 3.0 RAI manual allows for the entry of dashes in lieu of pain data. This is, therefore, a logical reason to exclude those residents for which no data are available. Analyses illustrate that missing data are weakly associated with higher scores on this measure, indicating that the use of dashes does not lead to better measure scores for facilities, which was a potential concern for the measure. Thus, not only do missing data do little to bias scores for this measure, they tend to be associated with slight *increases* in quality measure scores, suggesting that facilities with missing data on assessments have not seen any quality rating benefit from missing data.

Finally, for this measure, rates of missing data on pain items are not drastically different among individuals with selected characteristics related to pain (age, gender, cognitive impairment); however, these small differences are statistically significant.