NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

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

Brief Measure Information

NQF #: 0174

Corresponding Measures:

De.2. Measure Title: Improvement in bathing

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

De.3. Brief Description of Measure: Percentage of home health episodes of care during which the patient got better at bathing self.

1b.1. Developer Rationale: Patients need certain physical abilities and capacities to bathe themselves in the bath or shower. Many patients who receive home health care are recovering from an injury or illness and may have difficulty performing the tasks of bathing and/or may need help from

another person or special equipment to accomplish this activity. The required physical abilities for bathing can be developed or maintained by patient teaching or through rehabilitative services. Home health care staff can encourage patients to be as independent

as possible, can evaluate patients´ needs, and can teach them how to use special devices or equipment and increase their ability to perform some activities without the assistance of another person. Improving patients´ ability to bathe themselves contributes to patient

comfort, hygiene, skin integrity, quality of life and can allow them to live as long as possible in their own environment. Getting better at bathing may be a sign that they are meeting the goals of their care plan or that their health status is improving. Recovering independence in bathing is often a rehabilitative goal for home health patients, making it a reasonable evaluation indicator of effective and high-value home health care.

S.4. Numerator Statement: Number of home health episodes of care where the value recorded on the discharge assessment indicates less impairment in bathing at discharge than at start (or resumption) of care.

S.6. Denominator Statement: All home health episodes of care (except those defined in the denominator exclusions) in which the patient was eligible to improve in bathing (i.e., were not at the optimal level of health status according to the "Bathing" OASIS-C2 item M1830).

S.8. Denominator Exclusions: All home health episodes where at the start (or resumption) of care assessment the patient had minimal or no impairment, or the patient is non-responsive, or the episode of care ended in transfer to inpatient facility or death at home, or was covered by the generic exclusions.

De.1. Measure Type: Outcome

S.17. Data Source: Electronic Health Data

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Mar 31, 2009 **Most Recent Endorsement Date:** Jul 07, 2015

Preliminary Analysis: Maintenance of Endorsement

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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

Summary of prior evaluation in 2015

The developer cited clinical practice guidelines recommending functional assessment of older adults, including "independent performance of basic activities of daily living (ADLs), social activities, or instrumental activities of daily living (IADLs), the assistance needed to accomplish these tasks, and the sensory ability, cognition, and capacity to ambulate". The developer also provided literature linking home health interventions to improved outcomes, with three studies finding specific improvements in bathing associated with home health services. In the Friedman, et al (2014) article, authors linked improvements in bathing to teaching and support of patients and caregivers, environmental modifications, teaching use of assistive equipment, and strategies to mitigate associated pain and fatigue. NOTE: the 2015 NQF evidence criteria for outcome measures required a rationale to support the relationship of the health outcome to processes or structures of care.

The 2015 committee questioned the gap between the measured outcome and the evidence to support those interventions that would support improvement, noting there are no practice guidelines specifically around educating people on bathing. However, the Committee agreed that this is an important indicator because the goal is that home health patients to be independent and able to have

the ability to bathe themselves. Similar to all measures addressing improvements in ADLs, the Committee had a major concern about the requirement for CMS to not require improvement in function as a condition of coverage in home health, and applied the same remarks from the discussion on 0167 to all ADL improvement measures.

Changes to evidence from last evaluation

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

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

The developer provided additional literature, but other than calling out home-based occupational therapy targeted at physical exercise capacity of frail, older community-dwelling adults, no additional data was provided to to demonstrate the link between healthcare interventions and imrpvoement in bathing.

Question for the Committee:

- The evidence provided by the developer is updated and directionally the same compared to that for the previous NQF review. Does the Committee agree there is no need for repeat vote on Evidence?

Guidance from the Evidence Algorithm

Measure assesses a health outcome (Box 1) →The relationship between the outcome and the intervention demonstrated by performance data (Box 2) →Pass

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

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

Maintenance measures – increased emphasis on gap and variation

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

The developer provided performance data from June 2010 through 2016, which indicate room for improvement.

Risk Adjusted Home Health Agency (HHA) Level Performance on Improvement in Bathing by Calendar Year:

Calendar Year	Number of HHAs	Average Episodes per HHA	HHA Average	Std. Dev.	Min	25th %ile	50th %ile	75th %ile	Max	IQR*
HHAs with >=1 Valid Episode										
2010	10,861	294	59.6%	17.1%	0.0%	51.6%	61.9%	69.5%	100.0%	17.9%
2011	11,529	341	60.5%	16.8%	0.0%	52.5%	62.6%	70.3%	100.0%	17.8%
2012	11,791	332	61.4%	17.0%	0.0%	53.2%	63.7%	71.6%	100.0%	18.4%
2013	11,938	341	62.2%	17.5%	0.0%	53.6%	64.7%	72.7%	100.0%	19.1%
2014	11,877	359	62.2%	18.0%	0.0%	53.3%	65.0%	73.4%	100.0%	20.0%
2015	11,601	397	64.2%	18.5%	0.0%	55.5%	67.4%	75.8%	100.0%	20.3%
2016	11,221	404	67.6%	19.2%	0.0%	59.0%	71.4%	79.9%	100.0%	20.8%

Disparities

The developer provided data tables showing disparities in performance by race, age, gender, agency size, region, disability status and dual eligible status.

Observed and Predicted Episode-Level Measure Performance by Population Group:

Population Group		**2016 Observed**	**2016 Predicted**
All Episodes		74.5%	65.5%
Gender	Male	74.8%	66.3%
	Female	74.3%	64.9%
Race	White	75.6%	66.1%
	Black	72.0%	64.6%
	Hispanic	68.4%	61.2%
	Other	70.5%	62.4%
Age	Under 65	75.1%	66.7%
	65-74	80.3%	73.1%
	75-84	75.5%	66.4%
	85 and Over	66.9%	55.8%
Disability Status	No	74.8%	65.2%
	Yes	73.2%	66.2%
Dual Enrollment in Medicare and Medicaid	No	75.9%	66.3%
	Yes	69.5%	62.6%
Agency Size	Small	51.3%	56.3%
	Medium	68.9%	62.2%
	Large	75.9%	66.2%
Census Region	Northeast	75.0%	67.2%
	Midwest	73.7%	65.9%
	South	76.0%	65.4%
	West	71.5%	63.0%

Questions for the Committee:

- Does the measure demonstrate a quality problem related to home health care interventions and improvement in bathing?
- Is a national performance measure still warranted?
- Are you aware of evidence that other disparities exist in this area of healthcare?

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

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

1a. Evidence

Comments:

- Evidence shows direct relation between measured outcome and intervention.
- It seems that the evidence from 2015 has been strengthened by specific home-based OT targeted exercises in cited in the literature. I think this meets the criteria for sufficient evidence.
- Evidence describes home-based occupational therapy targeted at physical exercise capacity of frail, older community-dwelling adults but did not directly demonstrate the link between healthcare interventions and improvement in bathing. I am not aware of new studies that change the evidence base for this measure.
- The evidence is well documented. While some fit is tangential, the measure fits conceptually well with other evidence supporting the general maintenance of ADLs. Not aware of any new evidence.

- The developer has provided additional literature that called out home-based occupational therapy targeted at physical exercise capacity but additional data was not provided that demonstrated the link between healthcare interventions and improvement in bathing.
- The developer provided additional literature but it was not linked specifically to bathing. That said, there is a newly released study that does demonstrate a link between home health care and improvements in bathing: Rod Morgan & Rosanne DiZazzo-Miller (2019) The Occupation-Based Intervention of Bathing: Cases in Home Health Care, Occupational Therapy In Health Care, DOI: 10.1080/07380577.2018.1504368

1b. Performance Gap

Comments:

- Yes, there is still a high performance gap.
- The data provided demonstrates lots of opportunity for improvement.
- Data from 2010 to 2016 were presented noted significant room for improvement. Data showing subgroups was presented suggesting differences between groups, thought no statistical comparison was done. e.g. 68.4% v 75.6% for Hispanics and whites respectively
- While progress on the measure has been steady, the data clearly demonstrates that a performance gap clearly remains even in the top percentiles. Several disparities are also demonstrated.
- The data presented represents 2010 through 2016 and indicates a need for improvement. There are disparities in performance based on race, age, gender,disability status, agency size and region.
- Data through 2016 was provided and does demonstrate that there is definitely room for improvement. Data on the measure was provided on population subgroups and does demonstrate disparities. For example, compared to Whites, Blacks and especially Hispanic show a lower percentage in episodes of improved bathing upon discharge. Elders 85 and over also have a lower percentage than other age groups, as do those dually enrolled in Mediare and Medicaid and those receiving services from a small agency.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing

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

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

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

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

Evaluators: David Nerenz, Sam Simon, John Bott, Zhenqiu Lin, Joe Kunisch

Methods Panelists' Combined Preliminary Analysis

Evaluation of Reliability and Validity:

Reliability

- Reliability testing was conducted at the both the data element and measure score levels.
- Testing of the data elements
    - Developers conducted an inter-rater reliability (IRR) analysis among nurses and physical therapists using a linear weighted kappa statistic. Testing of OASIS-C2 item M1830 was done using 2016-2017 data from home health patients in 4 states.
        - Start Of Care/Resumption Of Care: kappa=0.51 (n=104 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
        - Discharge: kappa=0.43 (n=83 patients) ["Moderate" agreement, according to the Landis and Koch classification system]
- Testing of the measure score
    - Developers used two approaches to assess reliability of the measure score: a signal-to-noise analysis using the Adams beta-binomial method and a split-sample analysis using ICC(2,1) and ICC(3,1) statistics. CY2016 data were used in testing.
        - Signal-to-noise reliability estimates: Mean=0.93; minimum=0.64; 10^{th} percentile = 0.80; median =0.96; 90^{th} percentile =0.99
        - Split sample reliability estimates: IRR(2,1)= 0.89; IRR(3,1)= 0.89 [NOTE that testing data limited to agencies with ≥40 qualifying episodes]
        - Panel members would like to have seen data element validation for variables included in the risk-adjustment model (and any other critical data elements).

Validity

- Validity testing was conducted at the measure score level. The developer also described various data element validation assessments; however, results of these assessments were only summarized, not presented.
- Developers conducted a construct [convergent] validation analysis by correlating (using the Spearman's rank correlation coefficient) the results of this measure with 4 other OASIS performance measures (improvement in ambulation/locomotion, bed transfer, and pain interfering with activity, and management of oral medications) and a modified version of the Quality of Patient Care Star Rating measure (modified by excluding the bathing measure from the calculation).
    - Developers expected statistically significant, strong, positive correlations.
    - Correlations with the 4 OASIS measures ranged from 0.68-0.82.
    - Correlation with the modified star-rating measure = 0.76.
    - These results aligned with supported the developers' hypothesis.
- This measure is risk-adjusted using logistic regression with 120 risk factors (based on 2016 data).

- Developers discussed previous research linking dual-eligibility status and rural location with use of home health services. They therefore conducted analyses to examine associations between payment source (as a proxy for dual-eligibility) and rurality with this measure. They do include payment source in the risk-adjustment approach, but not rurality.
- Model discrimination:
  - Overall development sample: c-statistic=0.760
  - Overall model validation sample: c-statistic= 0.760
- Developers assessed risk-model calibration by calculating McFadden's R^2 and developing risk-decile plots.
  - Overall development sample: McFadden's R^2=0.152
  - Overall model validation sample: McFadden's R^2=0.147

- Panel members expressed some concern with excluding transferred patients, questioning whether those patients might have poorer outcomes on this measure. They had a similar concern with excluding patients who died.

Standing Committee Action Item(s):

- The Standing Committee can discuss reliability and/or validity, or accept the Scientific Methods Panel ratings.

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?
- The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

Preliminary rating for reliability:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**

Scientific Acceptability Preliminary Analysis

Measure Number: 0174
Measure Title: Improvement in Bathing

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**
☒ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☐ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**
☒ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☐ **Registry Data**
☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

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

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

☐ **Integrated Delivery System** ☐ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**
   Submission document: "MIF_xxxx" document, items S.1-S.22

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

   PANEL MEMBER 3: The OASIS items are well specified in the NQF evaluation form.
2. **Briefly summarize any concerns about the measure specifications.**
   PANEL MEMBER 1: None
   PANEL MEMBER 2: No concerns
   PANEL MEMBER 3: No concerns.
   PANEL MEMBER 4: One exclusion criterion does cause my concern. Episodes of care ended in transfer to in patient facility were excluded from the denominator because no assessment information was available for these patients. However, it is quite possible that many of these excluded patients might have poor outcomes. Given that a substantial proportion of episodes of care, about 27% (2b2.2), were excluded due to this reason, and particularly if there is across HHAs variation on this exclusion, the measure score may be potentially biased.
   PANEL MEMBER 5: None

RELIABILITY: TESTING

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

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

   PANEL MEMBER 3: Since data source testing was conducted, appears this question is not answered.
6. **Assess the method(s) used for reliability testing**
   Submission document: Testing attachment, section 2a2.2

   PANEL MEMBER 1: The methods used seemed adequate.
   PANEL MEMBER 2: The measure developer used appropriate methods to compute reliability estimates at the agency level using a STN model and conducted item-level analyses using kappa agreement to evaluate inter-rater reliability of assessment items

PANEL MEMBER 3: Measure score: "…fit a beta-binomial model to estimate measure reliability…" [p8]
"… test-retest reliability using the ICC to measure between-agency variation and within-agency variation…" [p9]
Data element: "…field test of new and existing OASIS items on 12 HHAs in four states for 213 home health patients. Home health registered nurses and physical therapists, trained by the study team, collected data during home visits at start of care (SOC) or resumption of care (ROC), and/or at discharge. Follow-up visits were conducted within 24 hours of the initial field test visit, by a different registered nurse or physical therapist to test interrater reliability…." [p10]
PANEL MEMBER 4: For reliability of the performance measure score, the developer tested both measure reliability (test – retest) and facility score reliability (beta-binomial). For the beta-binomial testing, however, it is not clear whether this testing was based on observed results or risk adjusted results. Because this measure is specified as a risk adjusted measure, the testing should be based on risk adjusted results.
PANEL MEMBER 5: The Kappa statistic was used to score the inter-rater reliability between the scores at first assessment for admitted or discharged patients. An independent trained team of RNs or physical therapist conducted a separate visit within 24 hours to independently assess the patient. 213 home visits were assessed across 4 states using the entire OASIS survey. For this measure, the underlying questions related to OASIS-C2 item M1830 (Bathing: Current ability to wash entire body safely)

7. **Assess the results of reliability testing**
Submission document: Testing attachment, section 2a2.3

PANEL MEMBER 1: Reliability seemed adequate – details below.

PANEL MEMBER 2: Results indicate both the assessment items and the agency-level scores are sufficiently reliable. Reliability of the bathing item at discharge was middling but acceptable.

While the raw outcome variable was shown to be reasonably reliable, I have 2 concerns. First, it was not clear whether the risk-adjusted score was modeled in the STN analysis. Given that the measure uses substantial risk adjustment, this seems like a significant omission. Further, item-level reliability was not reported for the 120 variables used in the risk –adjustment model, which also seems like important information to omit.

PANEL MEMBER 3: Measure score: "…fit a beta-binomial model to estimate measure reliability…" [p8]

"… test-retest reliability using the ICC to measure between-agency variation and within-agency variation…" [p9]

Data element: "…field test of new and existing OASIS items on 12 HHAs in four states for 213 home health patients. Home health registered nurses and physical therapists, trained by the study team, collected data during home visits at start of care (SOC) or resumption of care (ROC), and/or at discharge. Follow-up visits were conducted within 24 hours of the initial field test visit, by a different registered nurse or physical therapist to test interrater reliability…." [p10]

PANEL MEMBER 4: The summary of facility reliability scores (22a.3) showed that 75th percentile of facility score reliability is 0.99, indicating that 25% facility scores had reliability of 0.99, this is extremely rare for a risk adjusted measure. It is important to know if these results were based on unadjusted rates.
Weighted kappa is moderate, it would be helpful to report the proportions of agreement as well.

PANEL MEMBER 5: The mean and median inter-rater reliability scores of 0.93 and 0.96 for the entire OASIS survey, were above the range considered acceptable (0.70 – 0.80). Scores for the medication section were substantially lower indicated moderate agreement (0.51 and 0.43) at SOC/ROC.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2

   ☒**Yes**

   PANEL MEMBER 4: (ICC method is appropriate, beta-binomial approach is also appropriate but clarification needed for actual testing.)

   ☐**No**

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

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

   Submission document: Testing attachment, section 2a2.2

   ☒**Yes**

   ☒**No**

   PANEL MEMBER 2: results for risk adjustment variables were not included in item level or score-level testing

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

   PANEL MEMBER 3: Denominator exclusions in part include M1700, M1710, M1720. Ideally would have tested these data elements as well. Questionable whether these are 'critical' data elements, e.g. # / % of cases excluded by these OASIS questions.

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and <u>all</u> testing results):

    ☐**High** (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)

    ☒**Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)

    ☐**Low** (NOTE: Should rate <u>LOW</u> if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☒**Insufficient** (NOTE: Should rate <u>INSUFFICIENT</u> if you believe you do not have the information you need to make a rating decision)

    PANEL MEMBER 3: Excerpt from #7 above : "…The inter-rater reliability (weighted kappa) values for M1830 (Bathing: Current ability to wash entire body safely) was 0.51 at SOC/ROC and 0.43 at discharge."
    Also, see response to #9 above

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

    PANEL MEMBER 1: Methods and results seemed to support reliability of the measure at both data element and measure score levels. The level of agreement between raters for individual patients seems a little low, even though the developers claim that it is adequate. The generic adjective labels used to cover various ranges of the kappa statistic were developed in other contexts, and generally are based on a "null hypothesis" of no agreement at all. In a specific health care setting like home health care, and with a familiar concept in that setting, one might expect higher levels of agreement between two experts assessing the same patient at the same

time. The weak results in this area don't seem to translate into poor reliability or validity at the measure score level, though.

PANEL MEMBER 2: Would expect the STN analysis to include the risk adjusted score for each facility and the inter-rater reliability analysis to include the risk adjustment variables as well. Without this information, we cannot tell if the computed (i.e., risk adjusted) scores are reliable.
PANEL MEMBER 3: Sufficient testing performed for measure score & data element. Not able to rate 'high' due to response to #7 above: "…The inter-rater reliability (weighted kappa) values for M1830 (Bathing: Current ability to wash entire body safely) was 0.51 at SOC/ROC and 0.43 at discharge." Result is only modest.

Only potential concern noted re question #9 above: Denominator exclusions in part include M1700, M1710, M1720. Ideally would have tested these data elements as well. Questionable whether these are 'critical' data elements, e.g. # / % of cases excluded by these OASIS questions.

PANEL MEMBER 4: Although clarification is needed for the beta-binomial test results and weighted kappa is moderate, test – retest results do indicate high reliability.

PANEL MEMBER 5: Statistical testing was appropriate and thorough. Although I thought the lower rate of agreement being much lower than the overall entire form agreement was somewhat concerning.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**
Submission document: Testing attachment, section 2b2.

PANEL MEMBER 1: None – measure exclusions seemed appropriate and reasonable.

PANEL MEMBER 2: None

PANEL MEMBER 3: Per the completed NQF endorsement form – S.8. Denominator Exclusions: "… the episode of care ended in transfer to inpatient facility or death at home…". In measure specifications we want to avoid excluding cases that may reflect poor quality care. Of course, the quality of care is precisely what we're trying to measure. The concern is a portion of such cases excluded (noted above) may be due to poor quality. Thus, the entity essentially gets a pass on these cases.

PANEL MEMBER 4: One exclusion criterion does cause my concern. Episodes of care ended in transfer to in patient facility were excluded from the denominator because no assessment information was available for these patients. However, it is quite possible that many of these excluded patients might have poor outcomes. Given that a substantial proportion of episodes of care, about 27% (2b2.2), were excluded due to this reason, and particularly if there is across HHAs variation on this exclusion, the measure score may be potentially biased.

PANEL MEMBER 5: No concerns identified very thorough analysis.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**
Submission document: Testing attachment, section 2b4.

PANEL MEMBER 1: None – the developers did a thorough job of addressing the meaningful differences in performance issue.

PANEL MEMBER 2: No concerns.

PANEL MEMBER 3: No concern given the percentile distribution on p. 25

PANEL MEMBER 4: No

PANEL MEMBER 5: None

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
Submission document: Testing attachment, section 2b5.

PANEL MEMBER 1: N/A

PANEL MEMBER 2: N/A

PANEL MEMBER 3: No concern as it states the only data source is OASIS. It appears OASIS captures everything required for the measure, which includes the exclusions of transfer to inpatient facility or death at home.

15. **Please describe any concerns you have regarding missing data.**
Submission document: Testing attachment, section 2b6. .

PANEL MEMBER 1: None – missing data clearly make a difference, and this will be particularly true for agencies with a small number of episodes from which to generate a score.

PANEL MEMBER 2: The measure developer assures us there are 'minimal issues with missing data' as the system apparently rejects forms with missing data. Still, actual rates of missing data would be helpful.

PANEL MEMBER 3: Given the response to this question it appears there is no missing data: "There are minimal issues with missing data because the OASIS submission system rejects assessments with missing values. The provider must then resubmit the assessment. " Thus, no concerns.

PANEL MEMBER 4: No

PANEL MEMBER 5: As noted by submitters; missing data is rejected and requires resubmission

16. **Risk Adjustment**

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

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

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☒ Yes ☒ No ☐ Not applicable
PANEL MEMBER 4: (Payment source as proxy for Medicaid coverage)

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

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

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No
16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☐ No
16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No
16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☐ No
16d.5.Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☐ No

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

PANEL MEMBER 1: No significant concerns – the approach for both clinical and social risk factors seemed appropriate and thorough.

PANEL MEMBER 2: Overall, the model appears to perform well. However, with 120 risk factors, over-fitting is a real possibility even with this large a sample, but the developer does not address this concern (i.e., use of predicted R-sq).

PANEL MEMBER 3: "…The overall model development sample c-statistic is 0.760. The overall model validation sample c-statistic is 0.760… The overall model. development sample R^2 is 0.152. The overall model validation sample R^2 is 0.147…. The plot below shows that the predicted and observed values are similar and monotonically increasing with predicted probability, both of which indicate a well calibrated model. Additionally, we consider the R^2 statistics (included in response to 2b3.6) to be sufficient indicators of model fit.." [p22-23, figure: 23]

PANEL MEMBER 4: The adjusted rate is set to 100% if the calculated rate is higher than 100%, the adjusted rate is set to 0% if the calculated rate is lower than 0%.
The developer should articulate the rationale for this approach and report how many facilities were impacted by this approach.

PANEL MEMBER 5: Submitters included justification for the limited SES factors used in the model. Logistic regression model was used appropriately to measure the effect of the chosen elements and included only the statistically significant variables in the risk adjusted model.

VALIDITY: TESTING

17. **Validity testing level:** ☒ **Measure score** ☒ **Data element** ☒ **Both**
18. **Method of establishing validity of the measure score:**

    ☒ **Face validity**

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

    ☐ **N/A (score-level testing not conducted)**
19. **Assess the method(s) for establishing validity**
    Submission document: Testing attachment, section 2b2.2

    PANEL MEMBER 1: Methods were generally reasonable, including examining correlations between this measure and other accepted quality of care measures for home health agencies**.** **T**he data reported here support validity, and similar measures of the same concept generally have demonstrated validity**.**

    PANEL MEMBER 2: Spearman rank correlations with other ADL measures were computed for the measure score. Inter-rater agreement was used to evaluate item level validity. However, it is unclear if the correlations use risk-adjusted scores.
    Inter-rater reliability scores (weighted kappa) were used as a proxy to establish item-level validity scores for the OASIS item using different raters at 2 different points in time (paired assessments across raters were done within 24 hours). Conceptually, this approach is problematic. For validity testing, one expects a gold standard against which to compare – which assessment is the gold standard in this scenario? Consequently, I find this approach to evaluating item-level validity not compelling. However, score level validity results are appropriately computed.

    PANEL MEMBER 3: Measure score: "Convergent validity refers to the extent to which measures that are designed to assess the same construct are related to each other. To evaluate the convergent validity of the measure, Abt calculated the Spearman rank correlations of the *Improvement in Bathing* measure with other relevant measures, including the publicly-reported measures of home health quality derived from OASIS assessments." [p11]

"…reports the Spearman rank correlation of the *Improvement in Bathing* measure with a version of the Quality of Patient Care Star Rating, where *Improvement in Bathing* is excluded from the calculation of the star rating in order to avoid mechanical correlations." [p12]
Data element: Re OASIS: "updated and improved based on input from clinicians and technical experts", "published in the Federal Register for comment… and no objections or suggestions for revision have been noted regarding" [p11] "Validity testing included:

1) Consensus validity by expert researcher/clinical panels for outcome measurement and risk factor measurement

2) Consensus validity by expert clinical panels for patient assessment and care planning

3) Criterion or convergent/predictive validity for outcome measurement/risk factor measurement

4) Convergent/predictive validity.: case mix adjustment for payment

5) Validation by patient assessment and care planning" [p12]

PANEL MEMBER 4: analyses as outlined are reasonable

PANEL MEMBER 5: Spearman rank correlation and expert validity

20. **Assess the results(s) for establishing validity**
Submission document: Testing attachment, section 2b2.3

PANEL MEMBER 1: Validity results seem adequate
PANEL MEMBER 2: Overall, the results show that the raw measure score is valid given positive correlations with other similar measures of ADL function.
PANEL MEMBER 3: Measure score: 'Spearman rank: 0.68 - 0.82"[p13]
Data element:

1) Consensus validity: "recommended for measuring patient outcomes…"
2) Consensus validity: "recommended for inclusion…"
3) Criterion or convergent/predictive validity: "found to be related to other indicators of health status and patient outcomes…"
Note the topic heading here is "data element". The response is not in regard to data element level.
4) Convergent/predictive validity: "Case mix adjustment for payment: The item was tested and is used in the grouping algorithm that, in part, determines the per-episode payment to home health agencies for care provided under the Medicare home health benefit."
Note I don't think case mix adjustment for payment equates to case mix adjustment for risk of an outcome measure. The 3M APR analogy: 1 grouping for severity of illness as it relates to resource consumption, 1 grouping for risk of mortality.
5) Validation by patient assessment and care planning: "reported by practicing clinicians to be effective and useful…" [p13]
PANEL MEMBER 4: Testing results are acceptable.

PANEL MEMBER 5: Submitters demonstrated strong correlation using the Spearman rank correlation and also expert/ clinical panel validity

21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**
Submission document: Testing attachment, section 2b1.
☒**Yes**

☐**No**
☐**Not applicable** (score-level testing was not performed)

22. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?** *NOTE that data element validation from the literature is acceptable.*

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

☒**Yes**

☒**No**

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

PANEL MEMBER 2: Item level reliability findings (kappa agreement) do not provide information about the validity of all the data elements required to compute this measure.

PANEL MEMBER 3: See response to #22: Convergent/predictive validity: "Case mix adjustment for payment: The item was tested and is used in the grouping algorithm that, in part, determines the per-episode payment to home health agencies for care provided under the Medicare home health benefit."

Note I don't think case mix adjustment for payment equates to case mix adjustment for risk of an outcome measure. The 3M APR analogy: 1 grouping for severity of illness as it relates to resource consumption, 1 grouping for risk of mortality.

23. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

☒**High** (NOTE: Can be HIGH only if score-level testing has been conducted)

☒**Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

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

☐**Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level <u>is required</u>; if not conducted, should rate as INSUFFICIENT.)

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

PANEL MEMBER 1: There are no strong concerns about validity, but a rating of "high" validity would seem to require some more direct evidence of some defined quality of care process measures having a causal relationship with the outcome measure. If the measure indicates that some home health agencies are 'better" than others, what exactly is it that they are doing that is "better"? Then, to what extent does the outcome measure faithfully reflect those differences? A measure with "high" validity would have to be able to demonstrate something like what fraction of the observed variance in the measure score is associated with underlying differences in quality of care, and show that that fraction is large and significant.

In addition, CMS payment policies for home health agencies have incentivized provision of therapy services and enrollment of patients who can be expected to improve in a number of functional domains (e.g., patients who are post-hip-replacement surgery vs. patients with chronic medical or mental health condiions). For-profit and stand=alone agencies have generally served more of these "profitable" patients than have not-for-profit and hospital-based agencies. These considerations would lead to a bias across a range of "improvement" measures in favor of those agencies serving acute rehab patients vs. chronically ill patients. Presumably, the clinical risk adjustment model used here takes this issue into account, but it would be useful to know that agencies caring for patients who are not necessarily expected to improve at all in a number of functional domains, where the treatment goal is to prevent further decline, are not disadvantaged in a measure like this one.

PANEL MEMBER 2: Although not clear if risk adjusted score was used to determine measure score correlations and the approach to determining item-level validity was not sufficient, the outcome variable's correlations with other ADLs (presumably raw score) indicates this measure has sufficient validity at the score-level.

PANEL MEMBER 3: In general, performed well in testing. Noted as medium vs high due to:
[1] See response to #22: Convergent/predictive validity: "Case mix adjustment for payment: The item was tested and is used in the grouping algorithm that, in part, determines the per-episode payment to home health agencies for care provided under the Medicare home health benefit."
Note I don't think case mix adjustment for payment equates to case mix adjustment for risk of an outcome measure. The 3M APR analogy: 1 grouping for severity of illness as it relates to resource consumption, 1 grouping for risk of mortality.
[2] See response to #22: Note in response to #21 above, CMS notes they used convergent validity. However the results are not noted here.
[3] See response to #22 – specifically #3 under "data element" heading: Note the topic heading here is "data element". The response is not in regard to data element level.
PANEL MEMBER 4: This is a valid measure, the main concern I have is with the exclusion criterion that I mentioned earlier.

PANEL MEMBER 5: I did not have any concerns of the validity methods used. As noted in questions 21 & 22, the submitters demonstrated solid analysis for validity testing.

ADDITIONAL RECOMMENDATIONS

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

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

2a1. Reliability-Specifications

Comments:

- No concerns
- I appreciate the Scientific Methods Panel's evaluation and concur that there is moderate evidence of reliability and validity.
- I do not have concerns about the likelihood that this measure can be consistently implemented.
- No concerns.
- I am happy to accept the evaluation of the Scientific Methods Panel and have nothing to add.
- Reading through the evidence and reliability test results, I feel confident in the likelihood that this measure can be consistently implemented. The kappa values were very high for test-retest reliability of performance measure scores. Although the inter-rater reliability kappa stats were lower.

2a2. Reliability – Testing

Comments:

- No concerns
- No
- Testing shows kappa of 0.51 at start of care and 0.43 and discharge of care. Signal to noise with a mean of 0.93. § Split sample reliability estimates: IRR(2,1)= 0.89; IRR(3,1)= 0.8
- No concerns. Recommendation of the Scientific Advisory Panel can be accepted
- No.

- No concerns.

2b1. Validity –Testing

Comments:

- No concerns, but would like to know more about the specific exclusions and how they may have impacted the results and why they were excluded.
- No
- correlated with with 4 other OASIS performance measures (improvement in ambulation/locomotion, bed transfer, and pain interfering with activity, and management of oral medications) ranged from 0.68-0.82. These are not the same as bathing, but reasonably similar.
- No concerns
- No
- No concerns.

2b4-7. Threats to Validity

- same answer as above in 7.
- The concern about excluding transferred patients and patients who died seems reasonable and not a threat to validity.
- Measures improvement in bathing. wouldn't identify deterioration in bathing of those with minimal problems bathing at start of care. OASIS data are collected by the home health agency during the care episode and transmitted electronically to the CMS national OASIS repository--this minimizing concerns about missing data.
- No concerns
- I am happy to accept the evaluation of the Scientific Methods Panel and have nothing to add.
- Indications that the measure identifies meaningful differences about quality include findings that there is still room for improvement, and in terms of the interquartile range, there is a great deal of spread. Finally, the measure's performance when compared to other metrics shows that differences were statistically significant and in the expected direction. Missing data not an issue because the OASIS submission system does not allow providers to upload assessments with missing data.

2b2-3. Other Threats to Validity

Comments:

- yes conceptual relationship shown
- Yes, the risk adjustment is appropriate.
- Exclude people who have no trouble bathing as this measure only describes improvements. A risk adjustment strategy is described
- One concern as in the 167. While the exclusion of patients from the denominator whose episode of care ends with admission to a facility is understandable from an operational point of view, it does eliminate a population whose results might lower the numerator as well. It might be helpful for the developers to test at least a sample of this population to see how the results compare to the overall sample.
- I have no concerns regarding validity aside from those mentioned by the Scientific Methods Panel, regarding the exclusion of transferred and deceased patients. I do not believe that this exclusion constitutes a threat to the validity of the measure.
- Studies were reviewed that included SDS variables and not all were included in the risk adjustment because not all were available from CMS data. Also, race/ethnicity was not included due to literature citing that it is not a proxy for social risk. Not quite sure why the focus only on SDS variables... I am not an expert on risk-adjustment strategy but reading through it the results appear acceptable and the risk-adjustment strategy appropriate

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The data for this measure comes from the OASIS dataset. OASIS captures assessment information during the home health episode of care. Collection and transmission of OASIS is a requirement for the Medicare Home Health Conditions of Participation.

Questions for the Committee:

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

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility

Comments:

- no concerns
- Because the completion of OASIS data is required it makes assessment of bating feasible.
- Part of Oasis which is required for for the Medicare Home Health Conditions of Participation. No concerns about how the data collection strategy can be put into operational use.
- All data elements are routinely generated, and documented in electronic sources.
- I have no concerns regarding data collection. It appears to be easily captured without undue burden.
- All required data elements are available and in electronic form. I have no concerns.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

OR

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

Accountability program details

- The measure is used in the following:
  - Home Health Compare (public reporting)
  - Home Health Star Ratings (internal quality improvement). Agencies receive a "Outcome Quality Measure Report" that allows agencies to benchmark their performance against other agencies across the state and nationally, as well as their own performance from prior time periods.
  - ***It is not clear from the submission whether this measure is also included in the Home Health Quality Reporting Program (HHQRP) and Home Health Value-Based Purchasing (HHVBP) program.***

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- The developer reports that home health agencies obtain feedback on the measure via quarterly Quality of Patient Care Star Rating Provider Preview Reports. Agencies are able to review for errors or submit questions via email. Additionally, HHQRP training was conducted for agencies in 2017.
- While the developer did not summarize the feedback from home health agencies, they did note that no requests for modifications have been made.

Additional Feedback:

- No additional feedback has been provided.

Questions for the Committee:

- How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?
- How has the measure been vetted in real-world settings by those being measured or others?

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

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- The developer provided data to show improvement results. The developer also described improvement over time within population subgroups. They noted that the large improvements from from 2015 to 2016 likely were due to the introduction of several initiatives that incorporate this measure (the Quality of Patient Care (QoPC) Star Ratings, a composite of this

measure and several others that has been publicly reported on Home Health Compare since July 2015, and the Home Health Value Based Purchasing (HHVBP) program).

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- The developer stated that recent improvement in this measure has been relatively large compared to historical trends due to the implementation of two initiatives that involve this measure (the QoPC Star Ratings and HHVBP, beginning in 2015 and 2016, respectively).

Potential harms

- The developer did not indicate any potential harms or benefits from this measure.

***Questions for the Committee**:*

- Do the benefits of the measure outweigh any potential unintended consequences?

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

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

4a1. Use - Accountability and Transparency

Comments:

- Agree that the measure is in use and provides feedback and comparison for national and regional benchmarking and consumer info on quality. Not sure on opportunity for feedback from HHA. Would be helpful to have that information.
- The measure is publicly reported through Home Health Compare. Feedback has been considered when changes were incorporated.
- Publicly reported via Home Health Compare. Agencies can benchmark their results with other agencies via o Home Health Star Ratings. Quartery feedback via Quality of Patient Care Star Rating Provider Preview Reports. No data from these are presented but developers note that no requests for modifications were made.
- Data is readily available to users and feedback solicited.
- The measure is used in Home Health Compare, for public reporting, Home Health Star Ratings for internal quality improvement allowing agencies to compare their performance to one another. Although the developer did not summarize the feedback they have obtained feedback and noted that no requests for modifications have been made.
- The data are being publicly reported in numerous ways and used. Also, HHAs can submit feedback via a email box that has been set up. At this time, the HHAs have not requested any modifications to the measure.

4b1. Usability – Improvement

Comments:

- Yes, the benefits outweigh any potential harms
- I think the benefits outweigh the consequences by brining clinicians' attention to the measure for improved patient care.
- improvement over time within population subgroups are described. Since doesn't measure deterioration, deterioration could be missed. Otherwise no specific harms.
- No harms evident.

- Large improvements from 2015 to 2016 were reported by the developers most likely due to several initiatives that incorporate the measure. The developer did not indicate any potential harms or benefits from this measure.
- The results can definitely be used to further the goal of high-quality healthcare because the results make clear that there are disparities. HHAs can use the data to examine how their practices may contribute to disparities and how can they can make needed revisions. The developer did not report on any actual unintended consequences and I cannot think of any. I do believe that their are benefits to using the measure though.

Criterion 5: Related and Competing Measures

Related measures

- 2287: Functional Change: Change in Motor Score
- 2321: Functional Change: Change in Mobility Score
- 2632: Long-Term Care Hospital (LTCH) Functional Outcome Measure: Change in Mobility Among Patients Requiring Ventilator Support
- 2634: Inpatient Rehabilitation Facility (IRF) Functional Outcome Measure: Change in Mobility Score for Medical Rehabilitation Patients
- 2774: Functional Change: Change in Mobility Score for Skilled Nursing Facilities
- 2775: Functional Change: Change in Motor Score for Skilled Nursing Facilities
- 2776: Functional Change: Change in Motor Score in Long Term Acute Care Facilities
- 2778: Functional Change: Change in Mobility Score for Long Term Acute Care Facilities
- 2612: CARE: Improvement in Mobility
- 2613: CARE: Improvement in Self Care

Harmonization

- NQF may ask the Committee to make recommendations for combining or harmonizing measures.

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

5. Related and Competing Measures

Comments:

- ten listed - yes harmonization should be considered
- It seems that the listed measures are global--measuring multiple types of functioning. I think that this measure of one specific activity focuses care on it in a more targeted way.
- There are related measures. Measures are not harmonized.
- A couple of related measures but none which seem to directly compete.
- There appear to be several related measures evaluating mobility and self-care in several institutional settings.
- There are 10 related measures so it would be good for the committee to examine these measures and be prepared to discuss the possibility of harmonization.

Public and Member Comments

No NQF members have submitted support/non-support choices as of January 25, 2019.
No comments have been submitted as of January 25, 2019.

Developer Submission

Additional evaluations and submission materials attachments…

1. Evidence and Performance Gap – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

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

nqf_evidence_attachment_7.1--BATHE-jsr.docx

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

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

Yes

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0174

Measure Title: Improvement in bathing

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

Date of Submission:

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete EITHER 1a.2, 1a.3 or 1a.4 as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: a Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence b that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: c a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence b that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence b that the measured structure leads to a desired health outcome.
- Efficiency: d evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

a. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

b. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

c. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

d. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

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

Outcome

☒ Outcome: Improvement in Bathing

☐ Patient-reported outcome (PRO):

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

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

☐ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

Appropriate home health care interventions should improve the rates of patients showing improvement in bathing.

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

Not applicable

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

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

Patients need certain physical abilities and capacities to bathe themselves in the bath or shower. Many patients who receive home health care are recovering from an injury or illness and may have difficulty performing the tasks of bathing and/or may need help from another person or special equipment to accomplish this activity. The required physical abilities for bathing can be developed or maintained by patient teaching or through rehabilitative services. Home health care staff can encourage patients to be as independent as possible, can evaluate patients' needs, and can teach them how to use special devices or equipment and increase their ability to perform some activities without the assistance of another person. Improving patients' ability to bathe themselves contributes to patient comfort, hygiene, skin integrity, quality of life and can allow them to live as long as possible in their own environment. Getting better at bathing may be a sign that they are meeting the goals of their care plan or that their health status is improving. Recovering independence in bathing is often a rehabilitative goal for home health patients, making it a reasonable evaluation indicator of effective and high-value home health care.

Clinical assessment of patients' ability to bathe is important. Bathing is one of the basic activities of daily living (ADLs). The onset of difficulty with bathing is may be a precursor to further ADL disability[1] and has been identified as one factor in a model predicting mortality.[2] Home health staff interventions targeted at improving patients' ability to bathe may also help to reduce these risks.

Various conditions and symptoms are associated with ADL disability, including the ability to bathe. Chronic symptoms such as joint pain and back pain are highly prevalent among older adults, and associated with basic ADL disability.[3] Low physical activity and slowness, two frailty phenotype

---

[1] Golding-Day M, Whitehead P, Radford K, Walker M. Interventions to reduce dependency in bathing in community dwelling older adults: a systematic review. Syst Rev. 2017 Oct 11;6(1):198.

[2] Suemoto CK, Ueda P, Beltrán-Sánchez H, Lebrão ML, Duarte YA, Wong R, Danaei G. Development and Validation of a 10-Year Mortality Prediction Model: Meta-Analysis of Individual Participant Data From Five Cohorts of Older Adults in Developed and Developing Countries. J Gerontol A Biol Sci Med Sci. 2017 Mar 1;72(3):410-416.

[3] Henchoz Y, Büla C, Guessous I, Rodondi N, Goy R, Demont M, & Santos-Eggimann B. (2017). Chronic symptoms in a representative sample of community-dwelling older people: a cross-sectional study in Switzerland. BMJ Open, 7(1), e014485.

components, were significantly linked to difficulty with ADLs.[4] Frailty has also been identified as a predictor of poor recovery from disability.[5] Presence of depressive symptoms is another factor associated with difficulty meeting needs for daily care such as eating, or bathing and changing clothes.[6] Home-based occupational therapy targeted at physical exercise capacity of frail, older community-dwelling adults has been shown to be beneficial.[7] In addition to interventions aimed at improving the components of the bathing activity, interventions targeted to address conditions adversely impacting ADL ability may result in patients' improved ability to manage these daily needs, including bathing.

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

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

☒ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Assessment of physical function** **Kresevic DM** **2012** **Kresevic DM. Assessment of physical function. In: Boltz M, Capezuti E, Fulmer T, Zwicker D, editor(s). Evidence-based geriatric nursing protocols for best practice. 4th ed. New York (NY): Springer Publishing Company; 2012. p. 89-103.** **http://www.guideline.gov/content.aspx?id=43918&search=ambulation**

[4] Provencher V, Beland F, Demers L, Desrosiers J, Bier N, Avila-Funes JA, et, al. Are frailty components associated with disability in specific activities of daily living in community-dwelling older adults? A multicenter Canadian study. Archives of gerontology and geriatrics. 2017 Nov 1;73:187-94.

[5] Wu Wu C, Kim DH, Xue QL, Lee DSH, Varadhan R, Odden MC. Association of Frailty with Recovery from Disability among Community-Dwelling Older Adults: Results from Two Large U.S. Cohorts. J Gerontol A Biol Sci Med Sci. 2018 Apr 10. doi: 10.1093/gerona/gly080.

[6] Xiang Xiang X, An R, Heinemann A. Depression and Unmet Needs for Assistance With Daily Activities Among Community-Dwelling Older Adults. Gerontologist. 2018 May 8;58(3):428-437. doi: 10.1093/geront/gnw262.

[7] Liu C-J, Wen-Pin Chang, Megan C. Chang; Occupational Therapy Interventions to Improve Activities of Daily Living for Community-Dwelling Older Adults: A Systematic Review. Am J Occup Ther 2018;72(4):7204190060p1-7204190060p11. doi: 10.5014/ajot.2018.031252.

Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	Major Recommendations: Assessment Parameters • Comprehensive functional assessment of older adults includes independent performance of basic activities of daily living (ADLs), social activities, or instrumental activities of daily living (IADLs), the assistance needed to accomplish these tasks, and the sensory ability, cognition, and capacity to ambulate (Campbell et al., 2004 [Level I]; Doran et al., 2006 [Level VI]; Freedman, Martin, & Schoeni, 2002 [Level I]; Kane & Kane, 2000 [Level VI]; Katz et al., 1963 [Level I]; Lawton & Brody, 1969 [Level IV]; Lightbody & Baldwin, 2002 [Level VI]; McCusker, Kakuma, & Abrahamowicz, 2002 [Level I]; Tinetti & Ginter, 1988 [Level I]). o Basic ADLs (bathing, dressing, grooming, eating, continence, transferring) o IADLs (meal preparation, shopping, medication administration, housework, transportation, accounting) o Mobility (ambulation, pivoting) • Older adults may view their health in terms of how well they can function rather than in terms of disease alone. Strengths should be emphasized as well as needs for assistance (Depp & Jeste, 2006 [Level I]; Pearson, 2000 [Level VI]). • The clinician should document baseline functional status and recent or progressive declines in function (Graf, 2006 [Level V]). • Function should be assessed over time to validate capacity, decline, or progress (Applegate, Blass, & Franklin, 1990 [Level IV]; Callahan et al., 2002 [Level VI]; Kane & Kane, 2000 [Level VI]). • Standard instruments selected to assess function should be efficient to administer and easy to interpret. They should provide useful practical information for clinicians and should be incorporated into routine history taking and daily assessments (Kane & Kane, 2000 [Level VI]; Kresevic et al., 1998 [Level VI]) (see the "Availability of Companion Documents" field for tools). • Interdisciplinary communication regarding functional status, changes, and expected trajectory should be part of all care settings and should include the patient and family whenever possible (Counsell et al., 2000 [Level II]; Covinsky et al., 1998 [Level II]; Kresevic et al., 1998 [Level VI]; Landefeld et al., 1995 [Level II]).

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Grade assigned is indicated for each study is shown in 1a.4.2. Definitions - Levels of Evidence • Level I: Systematic reviews (integrative/meta-analyses/clinical practice guidelines based on systematic reviews) • Level II: Single experimental study (randomized controlled trials [RCTs]) • Level III: Quasi-experimental studies • Level IV: Non-experimental studies • Level V: Care report/program evaluation/narrative literature reviews • Level VI: Opinions of respected authorities/consensus panels
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

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

The evidence review provided support for the importance of assessing bathing in older people. There were multiple recommendations and the grade of the evidence ranged from Level 1 (systematic reviews) through Level 6 (expert opinion).

Multiple recommendations are made. The most relevant recommendation is the need for assessment:

Comprehensive functional assessment of older adults includes independent performance of basic activities of daily living (ADLs), social activities, or instrumental activities of daily living (IADLs), the assistance needed to accomplish these tasks, and the sensory ability, cognition, and capacity to ambulate (Campbell et al., 2004 [Level I]; Doran et al., 2006 [Level VI]; Freedman, Martin, & Schoeni, 2002 [Level I]; Kane & Kane, 2000 [Level VI]; Katz et al., 1963 [Level I]; Lawton & Brody, 1969 [Level IV]; Lightbody & Baldwin, 2002 [Level VI]; McCusker, Kakuma, & Abrahamowicz, 2002 [Level I]; Tinetti & Ginter, 1988 [Level I]).

The type of study designs included: five systematic reviews (Level I evidence); one non-experimental study (Level IV) and three using expert opinion (Level VI).

Benefits are implied but not described in the CPG.

No harms were identified.

As part of the literature review described below, additional sources of evidence were found that support both the need to assess functional abilities with community-dwelling older adults and more specifically, with those receiving home health care services. Five studies were found supporting the need to measure functional abilities for a home health care population, with three studies finding specific improvements in bathing associated with home health services. One study was a qualitative study on bathing in community-dwelling elders and was included because it reinforces the importance of bathing for this population.

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

A search of guideline.gov with the terms "bathing" and "home care" did not return any relevant guidelines. Search for only "bathing" returned one guideline that provides evidence on the importance of assessment of bathing in older people. The other guidelines were condition- or disease-specific (e.g. osteoarthritis or muscular dystrophy). Searching was done for a 5 year period.

PubMed and Google Scholar searches were performed using key word "Home health care" in combination with each of the following key words: "Bathing," "Functional Status," "Function" and "Activities of Daily Living." The search was limited to 2006 – present.

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

A: 1) Dudgeon, B.J., Hoffmann, J.M., Ciol, M.A., Shumway-Cook, A., Yorkston, K.M. & Chan, L. (2008). Managing activity difficulties at home: A survey of beneficiaries. *Arch Phys Med Rehab, 89*(7), 1256-1261. 2) This descriptive study used data from the 2004 Medicare Current Beneficiary Survey to examine prevalence of functional difficulties experienced by community-dwelling Medicare beneficiaries (n=14,483). 3) ADL and IADL difficulties were reported by 31.3% and 42.2% of beneficiaries, respectively, with impairment in bathing reported by 12% of respondents. For individuals reporting at least 1 ADL impairment, personal help in conjunction with assistive technologies were required most frequently for bathing (32%). 4) While findings were not limited to those beneficiaries receiving home health care services, they reinforced the prevalence of functional impairments in the population most likely to receive home health care services following acute care discharge and/or exacerbation of a chronic condition. This underscores the need to evaluate functional ability for these patients receiving home health care services.

B: 1) Leff, B., Burton, L., Mader, S.L., Naughton, B., Burl, J., Greenough, W.B., Guido, S. & Steinwachs, D. (2009). Comparison of functional outcomes associated with hospital at home care and traditional acute hospital care. *JAGS, 57*, 273-278. 2) This study compared outcomes for patients cared for in a Hospital at Home program, in which Medicare-certified home health agencies provided services to community-dwelling patients in lieu of extended hospitalization (n=72) vs. those receiving all treatment in acute care hospitals (n=47). Self-reported data on five ADLs and seven IADLs from recall one month prior to the

initial hospitalization and two weeks post-hospital admission. 3) The hospital at home care group experienced improvements in functional abilities that approached statistical significance (mean change = .39, SD = 3.13; p = .10) while the acute care hospital group declined in both ADLs and IADLs. 4) This study suggests that home health care services are associated with improvements in functional ability, including bathing.

C: 1) Scharpf, T.P. & Madigan, E.A. (2010). Functional status outcome measures in home health care patients with heart failure. *Home Health Services Quarterly, 29*(4), 155-170. 2) Data from OASIS ADL and IADL items were evaluated from a sample of 95,948 home healthcare patients with a diagnosis of heart failure. Changes over time in individual functional variables and in an index of ADLs were evaluated, along with patient variables that predicted improvement in the ADL index. 3) Bathing scores improved between home health admission and discharge (mean change score = -0.14, SD .25). ADL change scores reflecting improvement were predicted by worse ADL scores on admission, better oral medication scores at admission, age < 85, better cognitive function at admission, absence of urinary incontinence, and worse rehabilitation prognosis on admission. 4) The findings highlight the importance of measuring baseline functional status on multiple variables, including ambulation, and suggests that home health care services may facilitate improvements in functional ability, and bathing in particular, for heart failure patients.

D. 1) Friedman, B., Li, Y., Lievel, D.V. & Powers, B.A. (2014). Effects of a home visiting nurse intervention of care versus care as usual on individual activities of daily living: A secondary analysis of a randomized controlled trial. *BMC Geriatrics,* 14:24. Retrieved from http://www.biomedcentral.com/1471-2318/14/24. 2) A secondary analysis was conducted on data from a RCT of a home visiting intervention to facilitate chronic disease management in Medicare beneficiaries with significant functional impairment. The intervention consisted of monthly home visits by nurses implementing behavioral interventions to improve patient self-management. Impairment on six ADLS was compared at 22 months following study enrollment for the intervention (n = 384) vs. care as usual (n = 262) groups. 3) After risk adjustments for baseline characteristics, fewer patients in the intervention group reported some bathing difficulty (OR = 0.58; p ≤ .05) or great bathing difficulty (OR = .40; p ≤ .01) compared to the control group. 4) The findings suggest that for patients with bathing impairments, home health nursing services are associated with improvements. Authors noted that nurse interventions to improve bathing function included teaching and support of patients and caregivers, environmental modifications, teaching in use of assistive equipment, and strategies to mitigate associated pain and fatigue. Such interventions are consistent with evidence on effective strategies to minimize bathing disabilities.

E. 1) Ahluwalia, S.C., Gill, T.M., Baker, D.I. & Fried, T.R. (2010). Perspectives of older persons on bathing and bathing disability: A qualitative study. *J Am Geriatr Soc, 58*(3), 450-456. 2) Using a grounded theory framework, qualitative data was obtained from interviews of 23 community-dwelling elders (≥ 78 yo). 3) Three themes emerged from the qualitative analysis: a) Importance and personal significance of bathing, including perceptions of importance of cleanliness, social expectations, and view of bathing as a pleasurable experience; b) Variability in attitudes, preferences and sources of bathing assistance, particularly desires to bathe independently vs. feeling more secure with assistance; and c) Anticipation and response to bathing disability, in which participants described how they were preparing for or responding to functional impairments in bathing. 4) Findings from this qualitative study highlight the importance of bathing and issues associated with current or future bathing impairments for a population that frequently receives home health care services. While the study was not specific to patients receiving home health care services, it underscores the importance of home health care assessment, care planning and interventions designed to address and improve bathing functional abilities.

F. 1) Gitlin, L.N., Winter, L., Dennis, M.P., Corcoran, M., Schinfeld, S. & Hauck, W.W. (2006). A randomized trial of a multicomponent home intervention to reduce functional abilities in older adults. JAGS, 54, 809-816. 2) This study tested a home therapy intervention (OT and PT) designed to reduce functional impairments and promote self-management in 319 community-dwelling elders reporting impairment in at least one ADL. Participants were randomized into intervention vs. control groups (no home care) and interviewed at 6 and 12 months. 3) An ADL index, mobility/transferring index, an IADL index were calculated from patient report data. At 6 months, the intervention group (n = 154) reported less difficulty with ADLs and IADLS than the control group n = 146), with largest benefits occurring in toileting (P = .049, 95% CI = -.035 to 0.00)and bathing (P = .02, 95% CI = -0.52 to - 0.06). Mobility/transfer impairments were lower but nonsignificant in the intervention group. The magnitude of differences between groups on ADL and IADL impairment was similar for the 12-month timepoint. 4) This study supports the notion that home health care services can positively impact ADLs and IADLs, thus the importance of measurement of bathing and other ADL outcomes.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

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

Patients need certain physical abilities and capacities to bathe themselves in the bath or shower. Many patients who receive home health care are recovering from an injury or illness and may have difficulty performing the tasks of bathing and/or may need help from another person or special equipment to accomplish this activity. The required physical abilities for bathing can be developed or maintained by patient teaching or through rehabilitative services. Home health care staff can encourage patients to be as independent as possible, can evaluate patients´ needs, and can teach them how to use special devices or equipment and increase their ability to perform some activities without the assistance of another person. Improving patients´ ability to bathe themselves contributes to patient comfort, hygiene, skin integrity, quality of life and can allow them to live as long as possible in their own environment. Getting better at bathing may be a sign that they are meeting the goals of their care plan or that their health status is improving. Recovering independence in bathing is often a rehabilitative goal for home health patients, making it a reasonable evaluation indicator of effective and high-value home health care.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Tables 1 and 2, below, and in the "Importance to Report" attachment show observed and predicted measure performance, respectively, for calendar years 2010 through 2016, including the number of HHAs and the average number of episodes for HHAs. For each table, the top panel shows this information for all HHAs with at least one episode for which the measure is available. The bottom panel shows this information for HHAs with at least 20 episode for which the measure is available.

Table 1.Observed HHA-level Performance on Improvement in Bathing by Calendar Year

Calendar Year	Number of HHAs	Average Episodes per HHA	HHA Average	Std. Dev.	Minimum	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile	Maximum	IQR*
HHAs with >=1 Valid Episode												
2010	10,861	294	55.9%	20.7%	0.0%	27.1%	45.4%	59.4%	69.3%	77.9%	100.0%	23.9%
2011	11,529	341	56.6%	20.4%	0.0%	28.6%	46.2%	59.9%	70.0%	78.9%	100.0%	23.8%
2012	11,791	332	57.4%	20.9%	0.0%	28.6%	46.3%	60.9%	71.4%	80.0%	100.0%	25.2%
2013	11,938	341	58.2%	21.6%	0.0%	27.3%	46.7%	61.8%	72.6%	82.2%	100.0%	25.9%
2014	11,877	359	57.8%	22.2%	0.0%	25.0%	45.8%	62.0%	73.1%	82.4%	100.0%	27.2%
2015	11,601	397	59.6%	23.1%	0.0%	25.5%	47.8%	64.0%	75.8%	85.1%	100.0%	28.1%
2016	11,221	404	62.6%	23.9%	0.0%	25.5%	50.0%	68.2%	79.7%	88.0%	100.0%	29.7%
HHAs with >=20 Valid Episode												
2010	8,649	367	59.0%	15.9%	0.0%	37.5%	50.1%	61.1%	69.9%	77.0%	100.0%	19.7%
2011	9,605	408	59.3%	16.5%	0.0%	36.7%	50.0%	61.4%	70.4%	78.2%	100.0%	20.4%
2012	9,816	397	60.3%	17.0%	0.0%	36.8%	50.7%	62.5%	71.9%	79.6%	100.0%	21.2%
2013	9,921	409	61.1%	17.7%	0.0%	37.2%	51.2%	63.6%	73.0%	81.5%	100.0%	21.8%
2014	9,748	436	61.4%	18.1%	0.0%	36.3%	51.0%	64.1%	73.8%	82.1%	100.0%	22.9%
2015	9,571	480	63.5%	18.6%	0.0%	37.5%	53.5%	66.3%	76.5%	84.6%	100.0%	23.1%
2016	9,146	494	67.1%	18.8%	0.0%	41.2%	57.3%	70.6%	80.4%	87.7%	100.0%	23.1%

*The IQR (interquartile range) is a measure of variability. It is calculated by subtracting the 25th percentile value from the 75th percentile value.

Table 2. Risk Adjusted HHA-level Performance on Improvement in Bathing by Calendar Year

Calendar Year	Number of HHAs	Average Episodes per HHA	HHA Average	Std. Dev.	Minimum	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile	Maximum	IQR*
HHAs with >=1 Valid Episode												
2010	10,861	294	59.6%	17.1%	0.0%	36.7%	51.6%	61.9%	69.5%	77.4%	100.0%	17.9%
2011	11,529	341	60.5%	16.8%	0.0%	38.4%	52.5%	62.6%	70.3%	78.5%	100.0%	17.8%
2012	11,791	332	61.4%	17.0%	0.0%	38.7%	53.2%	63.7%	71.6%	79.9%	100.0%	18.4%
2013	11,938	341	62.2%	17.5%	0.0%	38.2%	53.6%	64.7%	72.7%	81.6%	100.0%	19.1%
2014	11,877	359	62.2%	18.0%	0.0%	37.5%	53.3%	65.0%	73.4%	82.0%	100.0%	20.0%
2015	11,601	397	64.2%	18.5%	0.0%	38.4%	55.5%	67.4%	75.8%	84.2%	100.0%	20.3%
2016	11,221	404	67.6%	19.2%	0.0%	39.9%	59.0%	71.4%	79.9%	88.2%	100.0%	20.8%
HHAs with >=20 Valid Episode												
2010	8,649	367	61.7%	13.1%	0.0%	45.2%	55.2%	62.9%	69.5%	76.0%	100.0%	14.3%
2011	9,605	408	62.2%	13.7%	0.0%	44.8%	55.3%	63.5%	70.3%	77.4%	100.0%	15.0%
2012	9,816	397	63.3%	13.9%	0.0%	45.5%	56.3%	64.5%	71.7%	79.0%	100.0%	15.3%
2013	9,921	409	64.1%	14.5%	0.0%	45.4%	56.8%	65.7%	72.8%	80.5%	100.0%	16.0%
2014	9,748	436	64.5%	14.7%	0.6%	45.3%	56.9%	66.3%	73.6%	81.0%	100.0%	16.7%
2015	9,571	480	66.8%	15.1%	0.0%	47.1%	59.4%	68.7%	76.1%	83.5%	100.0%	16.7%
2016	9,146	494	70.6%	15.3%	0.0%	51.2%	63.5%	72.7%	80.2%	87.5%	100.0%	16.6%

*The IQR (interquartile range) is a measure of variability. It is calculated by subtracting the 25th percentile value from the 75th percentile value.

Table 3 provides characteristics of all home health patients in 2016 for which this measure could be calculated.

Table 3. Patients Characteristics - All Patients in Measure Calculation, 2016

Population Group		**# of Patients**	**% of Patients**
Total		4,536,567	100.0%
Gender	Male	1,705,116	37.6%
	Female	2,831,451	62.4%
Race	White	3,491,480	77.0%
	Black	585,995	12.9%
	Hispanic	331,269	7.3%
	Other	127,823	2.8%
Age	Under 65	740,203	16.3%
	65-74	1,218,525	26.9%
	75-84	1,376,248	30.3%
	85 and Over	1,201,591	26.5%
Disability Status	No	3,573,242	78.8%
	Yes	963,325	21.2%
Dual Enrollment in Medicare and Medicaid	No	3,501,162	77.2%
	Yes	1,035,405	22.8%

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

See attachment, "Importance to Report" for a tabular presentation of these data.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Tables 4 and 5 in the "Importance to Report" attachment, as well as below, show observed and predicted measure performance for population groups, respectively. Measure performance improved from 2013 to 2016 for all population groups. For some population groups, performance gaps between subgroups also diminished over time.

For example, for gender the difference between observed measure performance for males and females in 2013 was 1.1 percentage points. This difference slightly decreased to 0.6 percentage points in 2016. The difference in measure performance between those aged 65 to 74 and the 85+ and under 65 age groups also decreased over time.

For some population groups, disparities did increase. For example, the difference in measure performance between white and Hispanic patients increased from 2.5 in 2013 percentage points to 7.2

percentage points in 2016. The difference in performance between small and large agencies also widened over time.

Table 4. Observed Episode-Level Measure Performance by Population Group

Population Group		2013	2014	2015	2016
All Episodes		67.0%	68.0%	70.5%	74.5%
Gender	Male	67.7%	68.6%	71.0%	74.8%
	Female	66.6%	67.6%	70.2%	74.3%
Race	White	67.9%	68.8%	71.5%	75.6%
	Black	63.8%	65.2%	67.9%	72.0%
	Hispanic	65.4%	65.3%	65.7%	68.4%
	Other	64.6%	65.3%	66.9%	70.5%
Age	Under 65	67.8%	69.1%	71.4%	75.1%
	65-74	74.2%	75.1%	76.9%	80.3%
	75-84	68.0%	69.0%	71.6%	75.5%
	85 and Over	58.4%	59.0%	62.2%	66.9%
Disability Status	No	67.2%	68.2%	70.9%	74.8%
	Yes	66.6%	67.3%	69.1%	73.2%
Dual Enrollment in Medicare and Medicaid	No	68.6%	69.5%	72.1%	75.9%
	Yes	62.7%	63.4%	65.3%	69.5%
Agency Size	Small	49.6%	48.5%	48.8%	51.3%
	Medium	63.4%	63.5%	65.5%	68.9%
	Large	68.3%	69.4%	71.9%	75.9%
Census Region	Northeast	67.2%	68.5%	71.2%	75.0%
	Midwest	67.4%	68.1%	70.6%	73.7%
	South	67.7%	68.7%	71.3%	76.0%
	West	64.7%	65.5%	67.8%	71.5%

Table 5. Predicted Episode-Level Measure Performance by Population Group

Population Group		2013	2014	2015	2016
All Episodes		65.8%	66.3%	64.5%	65.5%
Gender	Male	66.6%	67.1%	65.4%	66.3%
	Female	65.3%	65.8%	64.0%	64.9%
Race	White	66.3%	66.8%	65.1%	66.1%
	Black	64.9%	65.6%	63.7%	64.6%
	Hispanic	63.1%	63.2%	60.6%	61.2%
	Other	63.7%	63.8%	61.4%	62.4%
Age	Under 65	66.7%	67.3%	65.7%	66.7%
	65-74	73.2%	73.7%	72.2%	73.1%
	75-84	66.7%	67.2%	65.5%	66.4%
	85 and Over	56.8%	57.1%	54.9%	55.8%
Disability Status	No	65.5%	66.0%	64.3%	65.2%
	Yes	66.7%	67.2%	65.3%	66.2%

Population Group		2013	2014	2015	2016
Dual Enrollment in Medicare and Medicaid	No	66.6%	67.1%	65.3%	66.3%
	Yes	63.4%	63.9%	61.8%	62.6%
Agency Size	Small	58.8%	58.5%	56.1%	56.3%
	Medium	63.3%	63.5%	61.4%	62.2%
	Large	66.6%	67.1%	65.3%	66.2%
Census Region	Northeast	67.2%	67.7%	66.1%	67.2%
	Midwest	66.5%	66.7%	64.9%	65.9%
	South	65.8%	66.5%	64.6%	65.4%
	West	63.1%	63.6%	61.9%	63.0%

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

See 1.b4

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

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

Health and Functional Status : Change

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

Elderly, Populations at Risk : Individuals with multiple chronic conditions

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits/Home-Health-Quality-Measures.html

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

This is not an eMeasure **Attachment:**

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

Attachment **Attachment:** isc_mstr_-V2.21.1-_FINAL_08-15-2017_-_combined_worksheets-636686551475687631.xlsx

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

Attachment **Attachment:** OASIS-C2-AllItems-10-2016-636686552762843350.pdf

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

Clinician

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

No

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

Not Applicable

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

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

Number of home health episodes of care where the value recorded on the discharge assessment indicates less impairment in bathing at discharge than at start (or resumption) of care.

S.5. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

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

Number of home health episodes from the denominator in which the value recorded for the OASIS-C2 item M1830 (“Bathing”) on the discharge assessment is numerically less than the value recorded on the start (or resumption) of care assessment, indicating less impairment at discharge compared to start of care.

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

All home health episodes of care (except those defined in the denominator exclusions) in which the patient was eligible to improve in bathing (i.e., were not at the optimal level of health status according to the “Bathing” OASIS-C2 item M1830).

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

All home health episodes of care (except those defined in the denominator exclusions) in which the patient was eligible to improve in bathing (i.e., were not at the optimal level of health status according to the "Bathing" OASIS-C item M1830).

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

All home health episodes where at the start (or resumption) of care assessment the patient had minimal or no impairment, or the patient is non-responsive, or the episode of care ended in transfer to inpatient facility or death at home, or was covered by the generic exclusions.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

Home health episodes of care for which [1] at start/resumption of care OASIS item M1830 = 0, indicating the patient was able to bathe self independently; OR (2) at start/resumption of care, OASIS item M1700 "Cognitive Functioning" is 4, or M1710 "When Confused" is NA, or M1720 "When Anxious" is NA, indicating the patient is non-responsive; OR (3) The patient did not have a discharge assessment because the episode of care ended in transfer to inpatient facility or death at home; OR (4) All episodes covered by the generic exclusions:

a. Pediatric home health patients - less than 18 years of age as data are not collected for these patients.
b. Home health patients receiving maternity care only.
c. Home health clients receiving non-skilled care only.
d. Home health patients for which neither Medicare nor Medicaid are a payment source.
e. The episode of care does not end during the reporting period.
f. If the agency sample includes fewer than 20 episodes after all other patient-level exclusions are applied, or if the agency has been in operation less than six months, then the data is suppressed from public reporting on Home Health Compare.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*

Not applicable

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

Statistical risk model

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Higher score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

1. Define an episode of care (the unit of analysis): Data from matched pairs of OASIS assessments for each episode of care (start or resumption of care paired with a discharge or transfer to inpatient facility) are used to calculate individual patient outcome measures.

2. Identify target population: All episodes of care ending during a specified time interval (usually a period of twelve months), subject to generic and measure-specific exclusions.

Generic exclusions: Episodes of care ending in discharge due to death (M0100_ASSMT_REASON[2] = 08).

Measure specific exclusions: Episodes of care ending in transfer to inpatient facility (M0100_ASSMT_REASON[2] IN (06,07), patients who are comatose or non-responsive at start/resumption of care (M1700_COG_FUNCTION[1] = 04 OR M1710_WHEN_CONFUSED[1] = NA OR M1720_WHEN_ANXIOUS[1] = NA), and patients independent in bathing at start/resumption of care (M1830_CRNT_BATHG[1] = 00).

Cases meeting the target outcome are those where the patient is more independent in bathing at discharge than at start/resumption of care:

M1830_CRNT_BATHG[2] < M1830_CRNT_BATHG[1].

3. Aggregate the Data: The observed outcome measure value for each HHA is calculated as the percentage of cases meeting the target population (denominator) criteria that meet the target outcome (numerator) criteria.

4. Risk Adjustment: The expected probability for a patient is calculated using the following formula:

P(x)=1/(1+e^(-(a+?¦?b_i x_i ?)))

Where:

P(x) = predicted probability of achieving outcome x

a = constant parameter listed in the model documentation

bi = coefficient for risk factor i in the model documentation

xi = value of risk factor i for this patient. See the attached zipped risk adjustment file for detailed lists and specifications of risk factors.

Predicted probabilities for all patients included in the measure denominator are then averaged to derive an expected outcome value for the agency. This expected value is then used, together with the observed (unadjusted) outcome value and the expected value for the national population of home health agency patients for the same data collection period, to calculate a risk-adjusted outcome value for the home health agency. The formula for the adjusted value of the outcome measure is as follows:

X(A_ra)= X(A_obs)+ X(N_exp)-X(A_exp)

Where:

X(Ara) = Agency risk-adjusted outcome measure value

X(Aobs) = Agency observed outcome measure value

X(Aexp) = Agency expected outcome measure value

X(Nexp) = National expected outcome measure value

If the result of this calculation is a value greater than 100%, the adjusted value is set to 100%. Similarly, if the result is a negative number the adjusted value is set to zero.

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

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

Not Applicable

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

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

Not Applicable

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

If other, please describe in S.18.

Electronic Health Data

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

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

The measure is calculated based on the data obtained from the Home Health Outcome and Assessment Information Set (OASIS-C2), which is a statutorily required core standard assessment instrument that home health agencies integrate into their own patient-specific, comprehensive assessment to identify each patient's need for home care. The instrument is used to collect valid and reliable information for patient assessment, care planning, and service delivery in the home health setting, as well as for the home health quality assessment and performance improvement program. Home health agencies are required to collect OASIS data on all non-maternity Medicare/Medicaid patients, 18 or over, receiving skilled services. Data are collected at specific time points (admission, resumption of care after inpatient stay, recertification every 60 days that the patient remains in care, transfer, and at discharge). HH agencies are required to encode and transmit patient OASIS data to the OASIS repositories. Each HHA has on-line access to outcome and process measure reports based on their own OASIS data to the OASIS repositories. Each HHA has on-line access to outcome and process measure reports based on their own OASIS data submissions, as well as comparative state and national aggregate reports, case mix reports, and potentially avoidable event reports. CMS regularly collects OASIS data for storage in the national OASIS repository, and makes measures based on these data (including the Improvement in Bathing measure) available to consumers and to the general public through the Medicare Home Health Compare website.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

Available at measure-specific web page URL identified in S.1

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

Facility

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

Home Care

If other:

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

Not Applicable

2. Validity – See attached Measure Testing Submission Form

Testing_Form_Bathing_20180730.docx,RiskAdjustmentModel-636686559531201856.zip

2.1 For maintenance of endorsement

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

Yes

2.2 For maintenance of endorsement

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

Yes

2.3 For maintenance of endorsement

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

Yes - Updated information is included

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

Measure Number (*if previously endorsed*)**:** 0174
Measure Title: Improvement in Bathing
Date of Submission: 8/1/2018

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 (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). *Contact NQF staff if more pages are needed.*
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- 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 e 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 f 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; g

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). h

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; i,j 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 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

e. 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).

f. 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.

g. 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.

h. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

i. Risk factors that influence outcomes should not be specified as exclusions.

j. 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.

NOTE: ALL TESTING CONDUCTED IN THIS FORM RELY UPON MORE RECENT DATA AND AN UPDATED RISK ADJUSTMENT MODEL COMPARED TO THE PREVIOUS NQF SUBMISSION. WE DO NOT MARK ANY RESPONSES IN RED BECAUSE MOST RESPONSES WERE UPDATED.

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: Electronic Clinical Data	☒ other: Electronic Clinical Data

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*).

Home Health OASIS-C2

1.3. What are the dates of the data used in testing? January 1, 2016 to December 31, 2016

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:	☐ other:

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*)

To calculate the intra-class correlation (ICC) as part of reliability testing, the measure developer included Medicare-certified agencies with at least 40 home health quality episodes ending between January 1, 2016 and December 31, 2016 and meeting the measure denominator criteria.[8] There were 8,059 such agencies (71.8 percent of the 11,221 agencies with at least one quality episode meeting the measure denominator criteria ending during the same time period). The sample included all quality episodes at

[8] A minimum of 40 episodes is used instead of the 20 episode criteria for public reporting because the ICC requires splitting each HHA into two samples. To ensure that each sample has a 20 episode minimum, we use a 40 episode minimum for the HHA when evaluating test-retest reliability.

these agencies (4,488,363 in total) meeting the measure denominator criteria ending between January 1, 2016 to December 31, 2016.

To calculate the beta-binomial scores (as part of reliability testing) and conduct analyses related to validity testing and exclusions, the measure developer included Medicare-certified agencies with at least 20 home health quality episodes ending between January 1, 2016 and December 31, 2016 and meeting the measure denominator criteria. There were 9,146 such agencies (81.5 percent of the 11,221 agencies with at least one quality episode meeting the measure denominator criteria ending during the same time period). The sample included all quality episodes meeting the measure denominator criteria at these agencies (4,519,611 in total) ending between January 1, 2016 to December 31, 2016.

The risk adjustment model was developed using OASIS national repository data from assessments submitted between January 1, 2016 and December 31, 2016 (which included ~ 6.4 million episodes of care).

1.6. How many and which <u>patients</u> 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*)

The table below identifies the patients by population group used to calculate the intra-class correlation (ICC) as part of reliability testing. As noted in section 1.5, these are the patients represented in Medicare-certified agencies with at least 40 home health quality episodes ending between January 1, 2016 and December 31, 2016 and meeting the measure denominator criteria, the data represented 3,528,047 patients.

Number/Percentage of Patients Represented in HHAs with At Least 40 Valid Episodes, By Population Group

Population Group		**# of Patients**	**% of Patients**
Total		3,528,047	100%
Gender	Male	1,335,601	37.86%
	Female	2,192,446	62.14%
Race	White	2,743,725	77.77%
	Black	435,023	12.33%
	Hispanic	249,112	7.06%
	Other	100,187	2.84%
Age	Under 65	577,321	16.36%
	65-74	980,210	27.78%
	75-84	1,070,546	30.35%
	85 and Over	899,970	25.51%
Dual Enrollment in Medicare and Medicaid	No	2,793,692	79.19%
	Yes	734,355	20.81%
Currently or Originally Eligible for Medicare due to Disability	No	2,817,628	79.86%
	Yes	710,419	20.14%
Location of HHA by Census Region	Northeast	773,194	21.92%
	Midwest	740,553	20.99%
	South	1,375,509	38.99%
	West	620,995	17.60%
	Missing	17,796	0.50%

The table below identifies the patients by population group used to calculate the beta-binomial scores (as part of reliability testing) and conduct analyses related to validity testing and exclusions. As noted in section 1.5, these are the patients represented in Medicare-certified agencies with at least 20 home health quality episodes ending between January 1, 2016 and December 31, 2016 and meeting the measure denominator criteria, the data represented 3,563,226 patients.

Number/Percentage of Patients Represented in HHAs with At Least 20 Valid Episodes, By Population Group

Population Group		# of Patients	% of Patients
Total		3,563,226	100%
Gender	Male	1,348,528	37.85%
	Female	2,214,698	62.15%
Race	White	2,761,664	77.50%
	Black	444,018	12.46%
	Hispanic	255,448	7.17%
	Other	102,096	2.87%
Age	Under 65	585,337	16.43%
	65-74	989,673	27.77%
	75-84	1,080,475	30.32%
	85 and Over	907,741	25.48%
Dual Enrollment in Medicare and Medicaid	No	2,813,056	78.95%
	Yes	750,170	21.05%
Currently or Originally Eligible for Medicare due to Disability	No	2,841,947	79.76%
	Yes	721,279	20.24%
Location of HHA by Census Region	Northeast	775,067	21.75%
	Midwest	751,372	21.09%
	South	1,392,544	39.08%
	West	626,447	17.58%
	Missing	17,796	0.50%

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.

To calculate the intra-class correlation (ICC) as part of reliability testing, the measure developer included Medicare-certified agencies with at least 40 home health quality episodes ending January 1, 2016 and December 31, 2016 and meeting the measure denominator criteria.[1] There were 8,059 such agencies (71.8 percent of the 11,221 agencies with at least one quality episode meeting the measure denominator criteria ending during the same time period). The sample included all quality episodes at these agencies (4,488,363 in total) meeting the measure denominator criteria ending between January 1, 2016 to December 31, 2016.

To calculate the beta-binomial scores (as part of reliability testing) and conduct analyses related to validity testing and exclusions, the measure developer included Medicare-certified agencies with at least 20 home health quality episodes ending between January 1, 2016 to December 31, 2016 and meeting the measure denominator criteria. There were 9,146 such agencies (81.5 percent of the 11,221 agencies

with at least one quality episode meeting the measure denominator criteria ending during the same time period). The sample included all quality episodes meeting the measure denominator criteria at these agencies (4,519,611 in total) ending between January 1, 2016 to December 31, 2016.

The risk adjustment model was developed using OASIS national repository data from assessments submitted between January 1, 2016 and December 31, 2016 (which included ~ 6.4 million episodes of care).

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

We examined social risk factors that are available from the OASIS, as shown below. For operational and logistical reasons related to the monthly processing of this measure, drawing risk factors from outside external sources is not currently possible.

- Sex (female, male)
- Age (in 10 categories)
- Payment source (proxy for Medicaid coverage and dual eligibility using M0150 - Current Payment Sources for Home Care – see table below for the OASIS item responses).
  Response for M0150 – Current Payment Sources for Home Care (Mark all that apply)

M0150	Responses
0	None; no charge for current service
1	Medicare (traditional fee-for-service)
2	Medicare (HMO/managed care/Advantage plan)
3	Medicaid (traditional fee-for-service)
4	Medicaid (HMO/managed care)
5	Workers' compensation
6	Title programs (for example, Title III, V, or XX)
7	Other government (for example, TriCare, VA)
8	Private insurance
9	Private HMO/managed care
10	Self-pay
11	Other (specify)
UK	Unknown

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*)

Below, we address reliability at two levels: (1) the performance measure and (2) the underlying data element - OASIS item (M1830 Bathing: Current ability to wash entire body safely).

- **Reliability of the Performance Measure Score:** Abt measured the extent to which differences in each quality measure were due to actual differences in agency performance versus variation that arises from measurement error. Statistically, reliability depends on performance variation for a measure across agencies, the random variation in performance for a measure within an agency's panel of attributed beneficiaries, and the number of beneficiaries attributed to the agency. High reliability for a measure suggests that comparisons of relative performance across agencies are likely to be stable over different performance periods, and that the performance of one agency on the quality measure can confidently be distinguished from another. Potential reliability values range from zero to one, where one (highest possible reliability) means that all variation in the measure's rates is the result of variation in differences in performance across agencies, while zero (lowest possible reliability) means that all variation is a result of measurement error.

  Following the approach described by Adams,[9] Abt fit a beta-binomial model to estimate measure reliability. The beta-binomial model is appropriate because a particular agency's measure rate follows a binomial distribution (i.e., all measures are pass/fail), and it is reasonable to assume that the agencies' true measure rates vary and follow a beta distribution. It is reasonable to use the beta distribution to fit the true measure rates because it is a flexible distribution on the interval from 0 to 1, can have any mean on the interval, and can be skewed left, right, or U-shaped.

  Equation (1), which is based on the beta-binomial model, shows that reliability is dependent on two variance components: the variation across agencies, and variation within agencies. In general, reliability for agencies will be higher when the measure rates across agencies are more heterogeneous (as measured by the agency-to-agency variation). Agencies with larger samples (n) and pass rates (p) nearer to 0 or 1 will have higher levels of reliability because the agency-specific error is reduced (i.e. the estimated agency rates are more precise).

$$\text{Reliability} = \frac{\sigma^2_{\text{agency-to-agency}}}{\sigma^2_{\text{agency-to-agency}} + \sigma^2_{\text{agency-specific-error}}} = \frac{\sigma^2_{\text{agency-to-agency}}}{\sigma^2_{\text{agency-to-agency}} + \frac{p(1-p)}{n}} \quad (1)$$

  Abt also calculated the test-retest reliability using the ICC to measure between-agency variation and within-agency variation. First, we randomly divided home health episodes within each agency into two separate equally-sized groups. Then, we calculated performance rates for each group. Then, using the paired performance rates, we calculated the statistics absolute-agreement ICC (AA-ICC or ICC(2,1)) and consistency-of-agreement ICC (CA-ICC or ICC(3,1)) . ICC values that approach 1 indicate that the fraction of the total variance due to between-agency variation is high.

- **Reliability of the Underlying Data Element:** The measure is calculated by comparing patient functioning at the start and end of a home health quality episode, as reported by the home

[9] For more information about reliability testing for performance measurement, as well as the methodology for constructing the reliability score reported on Table 6, see "Reliability of Provider Profiling: A Tutorial" by John Adams, RAND. http://www.rand.org/pubs/technical_reports/TR653.html

health OASIS-C2 data set. Patient ability to ambulate is based on response to OASIS-C2 item M1830 (Bathing: Current ability to wash entire body safely):

> 0 - Able to bathe self in shower or tub independently, including getting in and out of tub/shower.
>
> 1 - With the use of devices, is able to bathe self in shower or tub independently, including getting in and out of the tub/shower.
>
> 2 - Able to bathe in shower or tub with the intermittent assistance of another person:
>
> (a) for intermittent supervision or encouragement or reminder, OR
>
> (b) to get in and out of the shower or tub, OR
>
> (c) for washing difficult to reach areas.
>
> 3 - Able to participate in bathing self in shower or tub, but requires presence of another person throughout the bath for assistance or supervision.
>
> 4 - Unable to use the shower or tube, but able to bathe self independently with or without the use of devices at the sink, in chair, or on commode.
>
> 5 - Unable to use the shower or tub, but able to participate in bathing self in bed, at the sink, in bedside chair, or on commode, with the assistance or supervision of another person.
>
> 6 - Unable to participate effectively in bathing and is bathed totally by another person.

In 2016 and 2017, Abt and partners conducted a field test of new and existing OASIS items on 12 HHAs in four states for 213 home health patients.[10] Home health registered nurses and physical therapists, trained by the study team, collected data during home visits at start of care (SOC) or resumption of care (ROC), and/or at discharge. Follow-up visits were conducted within 24 hours of the initial field test visit, by a different registered nurse or physical therapist to test interrater reliability. M1830 was one of the existing OASIS-C2 items that was tested. Interrater reliability was assessed for SOC or ROC and at Discharge with a linear weighted kappa. The number patients for which inter-rater reliability could be tested was 104 at SOC/ROC and 83 at discharge.

The kappa statistic is generally considered to be the "gold standard" statistic associated with item reliability as it factors in the possibility of chance agreement. Kappa values are reported as decimal values between 0.00 (poor) and 1.00 (perfect). These can be interpreted using the following seven categories:[11]

- Poor < 0.10
- Slight = 0.10 to 0.20
- Fair = 0.21 to 0.40
- Moderate = 0.41 to 0.60
- Substantial = 0.61 to 0.80
- Near perfect= 0.81 to 0.99
- Perfect = 1.00

[10] Abt Associates (2018). "OASIS Field Test Summary Report: Outcome and Assessment Information Set (OASIS) Quality Measure Development and Maintenance Project."

[11] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics, 1977. 33(1):159-174.

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*)

- **Reliability of the Performance Measure Score:** The table below summarizes the distribution of reliability scores for the 9,146 agencies that had at least 20 valid episodes.

Distribution of Beta Binomial Reliability Scores for Agencies with at Least 20 Valid Episodes

Mean	Minimum	10th Percentile	25th Percentile	Median	75th Percentile	90th Percentile	Maximum
0.93	0.64	0.80	0.90	0.96	0.99	0.99	1.00

For agencies with at least 40 valid episodes (recall that an ICC statistic is derived from paired performance rates), the AA-ICC is **0.887**, and the CA-ICC is also **0.887**.

- **Reliability of the Underlying Data Element:** The inter-rater reliability (weighted kappa) values for M1830 (Bathing: Current ability to wash entire body safely) was 0.51 at SOC/ROC and 0.43 at discharge.

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?*)

- **Reliability of the Performance Measure Score:** Using the beta-binomial model, Abt concluded that the measure reliability was high. The mean and median reliability scores of **0.93** and **0.96**, respectively, are above the range considered acceptable (0.70 – 0.80) for drawing inferences about home health agencies.

  The ICC statistics also suggest acceptable test-retest reliability.

- **Reliability of the Underlying Data Element** Based on the weighted kappa statistics the inter-rater reliability at SOC/ROC and at discharge was moderate (0.51 and 43, respectively). Given the scale of the response to this item (seven possible responses), we conclude that with moderate agreement, the item achieves sufficient reliability.

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)*

Below, we address validity at two levels: (1) the performance measure and (2) the underlying data element - OASIS item M1830 (Bathing: Current ability to wash entire body safely).

- **Validity of the Performance Measure Score:** Abt assessed the convergent validity of the measure. Convergent validity refers to the extent to which measures that are designed to assess the same construct are related to each other. To evaluate the convergent validity of the measure, Abt calculated the Spearman rank correlations of the *Improvement in Bathing* measure with other relevant measures, including the publicly-reported measures of home health quality derived from OASIS assessments.

  Abt also calculates and reports the Spearman rank correlation of the *Improvement in Bathing* measure with a version of the Quality of Patient Care Star Rating, where *Improvement in Bathing* is excluded from the calculation of the star rating in order to avoid mechanical correlations. The Spearman rank correlation assesses the statistical dependence between the rankings of two variables. In our case, we rank HHAs according to the *Improvement in Bathing* measure and other OASIS-based measures. High correlation or association between the *Improvement in Bathing* measure and other functional measures of improvement would be expected and desired. Low correlation would indicate that the measure may not be valid (is not measuring what we think it is measuring).

- **Validity of the Underlying Data Element:** The Bathing item has been used continuously as part of the OASIS since 2001. The behaviorally benchmarked responses were updated and improved based on input from clinicians and technical experts. The OASIS instrument has been published in the Federal Register for comment (both items and measures based off those items) and no objections or suggestions for revisits have been noted regarding the response options.
  The original OASIS item was originally carefully designed for measuring and ultimately enhancing patient outcomes as part of the National OBQI Demonstration project (1995 – 2000). OASIS items were derived by first specifying a set of patient outcomes considered critical by home care experts (e.g., nurses, physicians, therapists, social workers, administrators) for evaluating the effectiveness of care. These outcomes were chosen from the most important domains of health status addressed by home care providers. OASIS data items were developed, tested in hundreds of agencies, and refined for measuring outcomes in order to evaluate and enhance the effectiveness of home care. OASIS data items and measurement methods were reviewed by multidisciplinary panels of research methodologists, clinicians, home care managers, and policy analysts. Several tests of validity were conducted for each OASIS item, including Bathing. Validity testing included:

  1) Consensus validity by expert researcher/clinical panels for outcome measurement and risk factor measurement
  2) Consensus validity by expert clinical panels for patient assessment and care planning
  3) Criterion or convergent/predictive validity for outcome measurement/risk factor measurement
  4) Convergent/predictive validity: case mix adjustment for payment
  5) Validation by patient assessment and care planning

Descriptions for these validation assessments are taken from the "Volume 4 : OASIS Chronicle and Recommendation" OASIS and Outcome-Based Quality Improvement in Home Health Care, November 2001, Center for Health Services Research, University of Colorado Health Sciences Center, Denver, CO.

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

- **Validity of the Performance Measure Score:** The table below shows the Spearman rank correlations of the *Improvement in Bathing* measure with other publicly-reported measures of home health quality derived from OASIS assessments.

Spearman Rank Correlations of *Improvement in Bathing* Measure with Other Measures of Home Health Quality

Home Health Quality Measures	Spearman Rank Correlations
Improvement in Ambulation/Locomotion	0.8173
Improvement in Bed Transfer	0.6937
Improvement in Management of Oral Medications	0.6762
Improvement in Pain Interfering With Activity	0.6861
Quality of Patient Care Star Ratings (excluding *Improvement in Bathing*)	0.7590

- **Validity of the Underlying Data Element:** As noted above in 2b1.2,
  1. *Consensus validity:* The item was reviewed by panels of researchers and clinicians and was recommended for measuring patient outcomes relevant to home health care provision and quality measurement, or for risk adjustment of outcome analyses.
  2. *Consensus validity by expert clinical panels for patient assessment and care planning:* The item was reviewed by a panel of clinical experts and was recommended for inclusion in a core set of data items for patient assessment and care planning.
  3. *Criterion or convergent/predictive validity for outcome measurement/risk factor measurement:* The item was tested empirically for use in conjunction with outcome measures or risk factors predictive of patient outcomes. The item was found to be related to other indicators of health status and patient outcomes in a statistically significant and clinically meaningful way.
  4. *Convergent/predictive validity:* Case mix adjustment for payment: The item was tested and is used in the grouping algorithm that, in part, determines the per-episode payment to home health agencies for care provided under the Medicare home health benefit.
  5. *Validation by patient assessment and care planning:* The item has been used by clinicians for patient assessment and care planning in several hundred home health agencies and has been reported by practicing clinicians to be effective and useful for these purposes.

  Results of these validation assessments are taken from the "Volume 4 : OASIS Chronicle and Recommendation" OASIS and Outcome-Based Quality Improvement in Home Health Care, November 2001, Center for Health Services Research, University of Colorado Health Sciences Center, Denver, CO.

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?*)

- **Validity of the Performance Measure Score:** As detailed in the Spearman Rank Correlations table, the *Improvement in Bathing* measure displays a statistically significant positive correlation with several publicly-reported measures that similarly assess patient functioning and the quality of home health care, which lends evidence to the measure's validity. It may be that strong performance on the other measures directly leads to an improvement in bathing. It may also be the case that high quality agencies perform well on both the

Improvement in Bathing measure and other OASIS-based measures of patient functioning and communication due to cultural or organization-level factors.

- **Validity of the Underlying Data Element:** Item validity was established based on results of testing described in section 2b2.2, above. In addition, the item was also reviewed as part of the OMB/PRA review process for the most recent OASIS data set revision which allowed for two national comment periods (60 days and 30 days) wherein the face validity of the item was supported by the comments received.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b3

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*)

There are two major exclusion types for the *Improvement in Bathing* measure, including exclusions that are applicable to home health measures in general (i.e., generic exclusions) and exclusions that are specific to the *Improvement in Bathing* measure. Generic exclusions include (i) children and maternity patients and non-Medicare/non-Medicaid patients.

Exclusions that are specific to the *Improvement in Bathing* measure include (i) episodes of care that did not end in discharge to community, (ii) episodes were able to ambulate independently at baseline, and (iii) episodes in which the patient was non-responsive at baseline and therefore not expected to improve in bathing.

Abt calculated the frequency of the exclusions that are specific to the *Improvement in Bathing* measure, by exclusion type.

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*)

Measure Denominator Exclusion, January 2016 to December 2016

Home Health Stays	# of Episodes Excluded	% of Episodes Excluded	# of Episodes Remaining
A. All home health episodes	N/A	N/A	6,437,455
B. Home health episodes that exclude episodes that did not end in discharge to community	1,764,228	27.4	4,673,227
C. Home health episodes from B that exclude episodes for which the patient, at start/resumption of care, was able to ambulate independently	98,699	2.1	4,574,528
D. Home health episodes from C that exclude episodes in which the patient is nonresponsive	37,961	0.8	4,536,567

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*)

All the measure exclusions are conceptually justified, so the measure developer did not conduct further statistical analyses to test the exclusions. The remainder of this response provides justifications for the exclusions for the *Improvement in Bathing* measure.

Exclusions that are specific to the *Improvement in Bathing* measure include (i) episodes of care that did not end in discharge to community (i.e., episodes of care that ended in transfer to inpatient facility or death at home), (ii) episodes in which the patient was independent in bathing at baseline, and (iii) episodes in which the patient was non-responsive at baseline. For exclusion (i), the information needed to calculate the measure is not collected for these episodes of care. Exclusions (ii), and (iii) are justified because it would be impossible for these patients to demonstrate measurable improvement in bathing over the episode of care.

The generic exclusions for this measure include:

- *Children And Maternity Patients* - The OASIS data set items are designed to be collected for non-maternity, adult patients who are 18 years and older. Maternity patients, and patients less than 18 years of age are excluded.
- *Non-Medicare/non-Medicaid Patients* - Medicare-certified home health agencies are currently required to collect and submit OASIS data only on Medicare and Medicaid patients who are receiving skilled home health care.

If the agency sample includes fewer than 20 episodes after all other patient-level exclusions are applied, or if the agency has been in operation less than six months, then the data is suppressed from public reporting on Home Health Compare.

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.

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 120 risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

The *Improvement in Bathing* risk adjustment model includes 120 risk factors. The specification of the risk factors, estimated coefficients, and methodology are provided in the attachment.

2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

Not applicable; this 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?

The risk adjustment model was developed using OASIS national repository data from assessments submitted between January 1, 2016 and December 31, 2016 (~6.4 million episodes of care). The risk factors used in the unique prediction model created for each outcome measure are derived from OASIS data collected during the start of care or resumption of care assessment. The risk factors were developed and reviewed by home health clinicians. No ordering was used to determine risk factor inclusion, though, as described below, statistical criteria were applied to remove risk factors that were not statistically significant.

The risk adjustment methodology used is based on logistic regression analysis which results in a statistical prediction model for each outcome measure. For each home health agency patient who is included in the denominator of the outcome measure, the model is used to calculate the predicted probability that the patient will experience the outcome. The predicted probability for a patient is calculated using the following formula:

$$P(x) = 1/\left(1 + e^{-(a+\sum b_i x_i)}\right)$$

Where:

P(x) = predicted probability of achieving outcome x

a = constant parameter listed in the model documentation

b_i = coefficient for risk factor i in the model documentation

x_i = value of risk factor i for this patient

Predicted probabilities for all patients included in the measure denominator are then averaged to derive an expected outcome value for the agency. This expected value is then used, together with the observed (unadjusted) outcome value and the expected value for the national population of home health agency patients for the same data collection period, to calculate a risk-adjusted outcome value for the home health agency. The formula for the adjusted value of the outcome measure is as follows:

$$X(A_{ra}) = X(A_{obs}) + X\left(N_{exp}\right) - X(A_{exp})$$

Where:

$X(A_{ra})$ = Agency risk-adjusted outcome measure value

$X(A_{obs})$ = Agency observed outcome measure value

$X(A_{exp})$ = Agency expected outcome measure value

$X(N_{exp})$ = National expected outcome measure value

If the result of this calculation is a value greater than 100%, the adjusted value is set to 100%. Similarly, if the result is a negative number the adjusted value is set to zero.

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

☒ **Published literature**

☒ **Internal data analysis**

☐ **Other (please describe)**

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

We reviewed recent studies on accounting for sociodemographic status (SDS) conducted by the National Academies of Medicine (NAM), the Office of the Assistant Secretary for Planning and Evaluation (ASPE),

and NQF.[12] These studies tested SDS factors such as dual eligibility, rurality, race/ethnicity, and disability. While most of these variables are available via CMS data sources, we were not currently able to use other data sources to risk adjust this measure due to the operational requirements of producing this measure on a monthly basis. However, in the future, we plan to further investigate using the CMS Enrollment Database and other geographic-level files (such as the Area Health Resource File or Census data) to incorporate these other factors into the risk adjustment model.

We therefore were only able to include variables available on the OASIS. These include gender, payment source, age and race/ethnicity. We did not include race/ethnicity since it was not recommended as a proxy for social risk from the previous studies noted above. The payment source risk factor serves as a proxy for dual eligibility and Medicaid coverage. It tends to underreport dual eligibility and Medicaid coverage, but, as shown below, are important variables in explaining measure performance.

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.

We first present the observed measure performance between 2012 and 2016 stratified by each of the social risk factors. We note the greater increase in measure performance occurring between 2015 and 2016 (and to some extent between 2014 and 2015) may be related to the inception of the Quality of Patient Care Star Ratings and Home Health Value Based Purchasing. Both programs rely upon home health quality measures. The Quality of Patient Care Star Rating is a composite of a subset of measures reported on Home Health Compare, including *Improvement in Bathing*. The Home Health Value Based Purchasing program uses home health quality measures to generate a score that is compared across HHAs within a state (for nine states) and, depending on relative performance, can negatively or positively affect home health claims payment.

Differences in episode-level observed measure performance by gender were small, though, on average, males performed better on the measure than females in every year from 2012 to 2016.

Average Episode-Level Observed Measure Performance over Time, by Gender

	2012	2013	2014	2015	2016
Male	66.7%	67.7%	68.6%	71.0%	74.8%
Female	65.7%	66.6%	67.6%	70.2%	74.3%

Average episode-level observed measure values also differed by age group. Performance is concave with the youngest and oldest ages performing worse. Older patients performance substantially worse. Patients ages 65-70 performed best on the measure. These relationships were steady over time.

Average Episode-Level Observed Measure Performance over Time, by Age Category

[12] National Academies of Sciences, Engineering, and Medicine. (2016). Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. National Academies Press; Office of the Assistant Secretary for Planning and Evaluation (2016). Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. United States Department of Health and Human Services; National Quality Forum (2016). Early Results of SES Trial Reveal Need for Better Data and SES Variables. Available at: http://www.qualityforum.org/SES_Trial_Period_Update.aspx

	2012	2013	2014	2015	2016
0-54	66.3%	66.9%	68.4%	70.6%	74.1%
55-60	67.6%	68.5%	69.6%	72.1%	75.8%
60-65	67.5%	68.4%	69.7%	71.9%	75.8%
65-70	74.1%	75.2%	75.9%	77.7%	81.0%
70-75	72.2%	73.3%	74.3%	76.2%	79.7%
75-80	69.3%	70.2%	71.2%	73.6%	77.2%
80-85	65.1%	66.1%	66.9%	69.7%	73.9%
85-90	60.3%	61.3%	62.1%	65.3%	69.8%
90-95	54.7%	55.9%	56.5%	59.9%	64.8%
95+	47.9%	49.0%	49.8%	53.1%	58.5%

Average episode-level observed measure values were lower for patients using Medicare and Medicaid or Medicaid only as a payment source. Patients who indicated Medicare only performed the best on the measure.

Average Episode-Level Observed Measure Performance over Time, by Payment Source

	2012	2013	2014	2015	2016
Medicare and Medicaid	64.9%	65.9%	66.8%	63.3%	67.3%
Medicaid only	62.3%	62.3%	65.9%	68.9%	72.8%
Medicare only	66.6%	67.7%	68.6%	70.8%	74.8%

The following table displays the relevant estimated coefficients from the logistic regression model of *Improvement in Bathing* on a full set of OASIS-based risk factors (see Section 2.3.1.1). This table shows that male patients, patients who are age 65-69, and patients for whom the payer source is Medicare FFS are more likely to perform better on this measure. Almost all risk factors are statistically significant at the 1 percent statistical level.

	Coefficient	p-value
Female (excluded category)		
Male	0.059	0.000
AGE_0_54	-0.072	0.000
AGE_55_59	-0.084	0.000
AGE_60_64	-0.102	0.000
AGE_65_69 (excluded category)		
AGE_70_74	-0.018	0.014
AGE_75_79	-0.072	0.000
AGE_80_84	-0.157	0.000
AGE_85_89	-0.277	0.000
AGE_90_94	-0.455	0.000
AGE_95PLUS	-0.657	0.000
PAY_MCAID_ONLY	-0.224	0.000

	Coefficient	p-value
PAY_MCARE _FFS (excluded category		
PAY_MCAREANDMCAID	-0.372	0.000
PAY_MCARE_HMO	-0.112	0.000
PAY_OTHER	-0.063	0.002

To address the second part of this question – regarding the impacts of not adjusting for certain social risk factors for providers at extreme levels of risk, we take a closer look at the HHA's geographic locations – specifically, we compare observed to risk adjusted measure values for HHAs located in rural versus urban settings. Rural residents may have worse health outcomes and experience reduced access to health services, affecting their ability to improve on this measure.[13] The table below shows observed and risk adjusted measure values over time for rural and urban HHAs using a CBSA-based designation provided in the Provider of Services file

Risk-Adjusted and Observed Improvement in Bathing Measure Values by Rural and Urban Designation

		2011	2012	2013	2014	2015	2016
Observed	Rural	61.4%	62.6%	62.2%	62.8%	65.3%	69.3%
	Urban	58.8%	59.8%	60.9%	61.1%	63.1%	66.6%
Risk Adjusted	Rural	61.2%	62.9%	63.2%	63.8%	66.1%	70.2%
	Urban	62.4%	63.3%	64.3%	64.7%	66.9%	70.7%

This table shows that the differences between rural and urban HHAs were small and observed value for rural HHAs actually tended to be higher than urban HHAs. Differences in risk adjusted average scores remained very small on average, though urban HHAs tended to do slightly better than rural HHAs on average. Risk adjustment for rural and urban status is unlikely to change measure performance (in fact, we did not find the estimated coefficient on rural status to be statistically difference from zero when tested). As mentioned above, the greater increase in measure performance occurring between 2015 and 2016 (and to some extent between 2014 and 2015) may be related to the inception of the Quality of Patient Care Star Ratings and Home Health Value Based Purchasing.

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*)

Using the assessment data from January 1, 2016 to December 31, 2016, nearly 6.4 million episodes of care were created. This was done by linking the start of care (SOC) or resumption of care (ROC) assessment for a patient with that patient's last assessment (i.e., transfer, discharge, or death). We split the population of 6.4 million episodes for calendar year 2016 in half such that 3.2 million episodes were used as a developmental sample and 3.2 million episodes were used as a validation sample. A structured

[13] Befort, C. A., Nazir, N., & Perri, M. G. (2012). Prevalence of obesity among adults from rural and urban areas of the United States: findings from NHANES (2005-2008). *The Journal of Rural Health*, *28*(4), 392-397.

Dye, C., Willoughby, D., Aybar-Damali, B., Grady, C., Oran, R., & Knudson, A. (2018). Improving Chronic Disease Self-Management by Older Home Health Patients through Community Health Coaching. *International journal of environmental research and public health*, *15*(4), 660.

approach was used to develop the initial prediction model. The risk factors used in the prediction models are derived from OASIS data collected during the start of care or resumption of care assessment. Because there were a large number of possible risk factors that needed to be considered for the measure outcome and because some of the risk factors used previously are expected to be removed as part of the transition to OASIS-D in January 2019, the following process was used to identify unique contributing risk factors to the prediction model:

1. We identified risk factors based on OASIS items that will remain following the OASIS-D transition. We examine the statistical properties of the items to specify risk factors (e.g., we grouped item response when there was low prevalence of certain responses). Team clinicians then reviewed all risk factors for clinical relevance and we re-defined or updated risk factors as necessary. We then divided these risk factors into 35 content focus groups (e.g., ICD9-based conditions). Where possible, we defined risk factors such that they flagged mutually exclusive subgroups within each content focus group. When modelling these risk factors, we use the risk factor flag indicating independence as our exclusion category.

2. We use a logistic regression specification to estimate coefficients among the full set of candidate risk factors. Those risk factors that are statistically significant at probability <0.001 are kept for further review.

5. The list of risk factors that achieved the probability<0.01 level were reviewed. If one response option level of an OASIS-D item was on the list, then risk factors representing the other response option levels of that OASIS-D item were added to the list. For example, if response option levels 1 and 2 for M1800 Grooming were statistically significant at probability<0.01 for a particular outcome, then response option level 3 for M1800 Grooming was added to the list.

6. A fixed logistic regression was computed on the list of risk factors that had achieved probability<0.001 and the risk factors that were added to the list because they were other response options for OASIS-D items represented on the list.

7. Goodness of fit statistics (R^2 and c-statistic) as well as bivariate correlations between the risk factor and the outcome were computed for how well the predicted values generated by the prediction model were related to the actual outcomes.

8. The initial model was reviewed by a team of at least three experienced home health clinicians. Each risk factor was reviewed for its clinical plausibility in being related to the outcome measure in the direction indicated by the coefficient in the prediction equation and its bivariate relationship. Risk factors that were not clinically plausible were identified for elimination.

9. The risk factors that were deemed not clinically plausible were removed from the prediction model and steps 6 and 7 in this process were repeated. The resulting logistic regression equation was designated as the prediction model for the outcome.

10. The prediction model was applied to the validation sample and goodness of fit statistics were computed. If these statistics were similar to the goodness of fit statistics computed with the development sample, the model become a final model. If the statistics were not similar, then alternative approaches to model building were considered.

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

If stratified, skip to 2b3.9

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

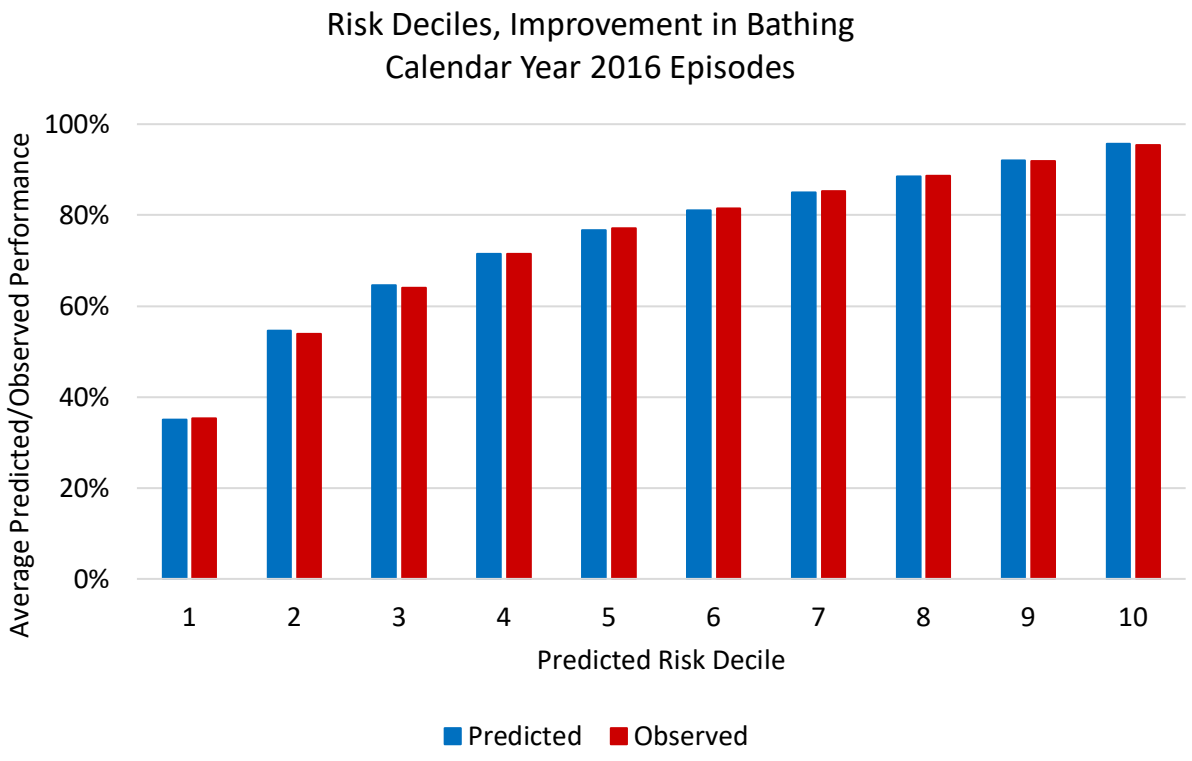
The c-statistic is the area under the Receiver Operating Characteristic curve. Intuitively, it is defined as follows: Let Y=1 denote outcome attainment, Y=0 denote nonattainment, and $\hat{p}$ denote the predicted probability that Y=1. Enumerate all possible pairs of sample patients for whom Y=1 for the first patient and Y=0 for the second patient. C is the proportion of such pairs where $\hat{p}$ for the patient with Y=1 is larger than $\hat{p}$ or the patient with Y=0. The overall model development sample c-statistic is **0.760**. The overall model validation sample c-statistic is **0.760.**

Because the risk adjustment model uses a logistic specification, we report McFadden's R^2 to summarize model fit. The traditional R^2 value for linear specifications is the squared correlation between predicted and observed values for all patients in the developmental or validation samples. McFadden's R^2 is conceptually similar and compares the likelihood the full model to an intercept-only model. The overall model development sample R^2 is **0.152**. The overall model validation sample R^2 is **0.147**.

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

With a validation sample of over 2 million episodes, the Hosmer-Lemeshow test will reject the null assumption of equality even if differences in average performance are small. As such, we prefer a visual inspection of the risk decile plot below, which compares the average predicted performance against the average observed performance for *Improvement in Bathing*. The plot below shows that the predicted and observed values are similar and monotonically increasing with predicted probability, both of which indicate a well calibrated model. Additionally, we consider the R^2 statistics (included in response to **2b3.6**) to be sufficient indicators of model fit.

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



2b3.9. Results of Risk Stratification Analysis:

Not applicable.

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*)

The c-statistic for the development sample is **0.760**, which is similar to the validation sample value of **0.760**, showing that the model differentiates between outcomes as well on new data as it does on the development data.

The McFadden's R^2 for the development sample is **0.152**, which is similar to the validation sample value of **0.147**, showing that the model is capable of describing the relationship between the covariates and the outcome in the development data set while also successfully predicting the outcome on a new data set.

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

None

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

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

To demonstrate that the *Improvement in Bathing* measure exhibits variation and that the variation is meaningful in discriminating performance among home health agencies, we conducted the following analyses:

1. First, we show that there is variation in the measure by examining the measure distribution – mean, median, 10th, 25th, 75th, and 90th percentile values. We also calculated the truncated coefficient of variation (TCV).
    a. We show that the measure is not "topped-out;" that is, we show there is room for improvement in the measure. Measures that are "topped-out" or close to being so are less able to meaningfully discriminate between providers. That is, if the majority of agencies are already performing at a high level, the measure is less able to distinguish between providers. We demonstrate that the 10th percentile value of the measure is less than 70 percent. That is, if the HHAs performing at the 10th percentile had a measure value of 70 percent, then we would consider the measure having little room for improvement.
    b. We show that the interquartile range (IQR) is substantial. The IQR is calculated by subtracting the 25th percentile measure value from the 75th percentile measure; it shows the measure "spread."
    c. The TCV is another measure of variation – it is the ratio of the truncated standard deviation and truncated mean. We truncate by removing the bottom 5th percentile and the top 95th percentile of HHAs. A larger TCV indicates higher variability of the measure.
    d. We show the same information for HHA stratified by whether the census region in which the HHA is located.
2. Demonstrating that there is variation in the measure is not sufficient for concluding that the variation is meaningful. To examine whether the measure is meaningful in distinguishing performance across agencies, we examined the performance of the measure by an altered

version of the Quality of Patient Care (QoPC) Star Rating and tested whether measure values differ by rating and whether the difference is statistically significant at the 5 percent significance level. The QoPC Star Rating is composed of eight equally weighted quality measures, including Improvement in Bathing.[14] We created an altered version that removes the *Improvement in Bathing* from the QoPC Star Ratings (keeping the remaining measures and methodology the same). The other measures include other functional improvement measures, two process measures and a claims-based hospitalization measure. The QoPC Star Ratings are a composite of these measures and take on nine values (1 to 5 stars in half star increments). Higher stars indicate higher quality. We thus expect that HHAs with higher QoPC Star Ratings (or alternate) values will have higher values on the *Improvement in Bathing* measure.

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*)

The table below shows the distribution of the *Improvement in Bathing* measure across the 9,146 agencies that had at least 20 episodes available. The median is 68.2 percent. The 10th percentile value is 25.5 percent and the 90th percentile value is 88.0 percent. The IQR is 29.7 percent. The TCV (not shown in the table) is 38.2 percent. These statistics show that the measure is not topped out and there is still sufficient room for improvement.

Distribution of Improvement in Bathing (Risk Adjusted) Overall and by Census Region

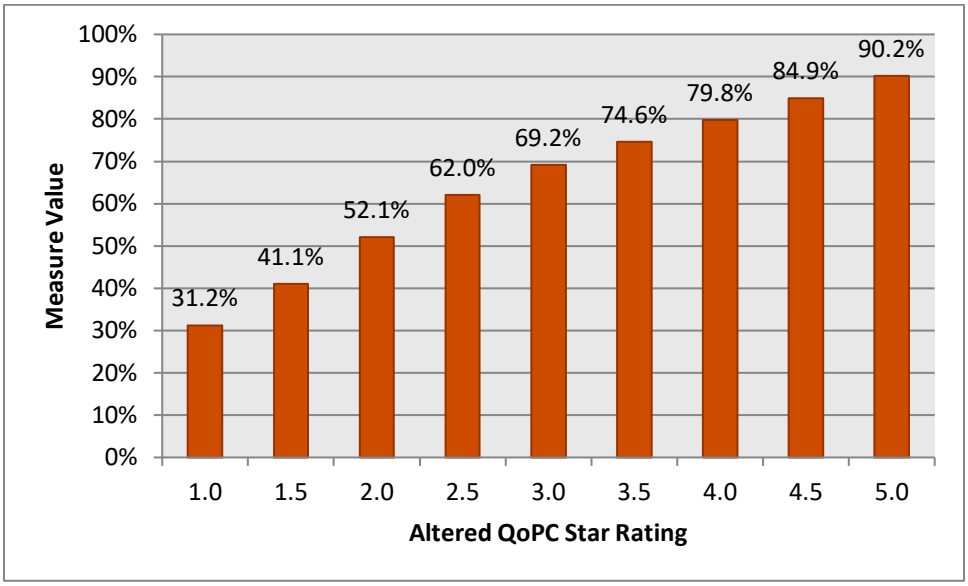
	#HHAs	Mean	10th	25th	50th	75th	90th	IQR
All*	9,146	62.6%	25.5%	50.0%	68.2%	79.7%	88.0%	29.7%
Northeast	776	62.1%	25.0%	50.2%	69.8%	78.7%	85.4%	28.4%
Midwest	2,428	60.8%	23.1%	49.2%	66.1%	78.6%	87.0%	29.4%
South	4,072	62.0%	23.5%	47.8%	68.4%	80.5%	89.0%	32.7%
West	1,829	66.8%	42.5%	57.5%	69.6%	79.5%	88.5%	22.0%

*Note that "All" includes all HHAs in the 50 states and U.S. territories. The census regions only include U.S. States (thus, the number of HHAs in each census region does not all up to "All").

This figure and table below shows the measure value by "altered" QoPC Star Rating. The figure shows that the *Improvement in Bathing* measure steadily increases with a higher rating. The table below the figure shows the same information in table format. It includes the count of the number of HHAs with each rating as well as the statistical significance of a t-test between with sequential pairing. For example a t-test of the difference between the measure value for HHAs with 1.0 stars versus HHAs with 1.5 stars showed that the difference was different from zero with a p-value of 0.000 (i.e., statistically significant at the 5 percent level). All sequential pairwise differences were statistically significantly different from zero.

[14] https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits/Downloads/QoPC-Methodology_for_April_2018.pdf

Measure Performance by "Altered" Quality of Patient Care Star Rating*



Altered QoPC Star Rating	HHA Count	Risk Adjusted Measure Value	Pairwise p-value
1.0	31	31.2%	-
1.5	247	41.1%	0.000
2.0	828	52.1%	0.000
2.5	1,371	62.0%	0.000
3.0	1,808	69.2%	0.000
3.5	1,886	74.6%	0.000
4.0	1,517	79.8%	0.000
4.5	946	84.9%	0.000
5.0	328	90.2%	0.000
Missing	184	62.9%	-

*The QoPC Star Rating was altered by removing the *Improvement in Bathing* measure from the rating calculation.

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?*)

Based on these findings, we conclude that the *Improvement in Bathing* measure is able to produce meaningful differences across HHAs. First, the measure exhibits sufficient variation – it is not topped out and there is room for measure improvement among the majority of HHAs. Second, measure performance is related to other metrics in the direction expected with statistically significant differences in measure performance across strata.

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*)

N/A – one set of data/specifications are used

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A – one set of data/specifications are used

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*)

N/A – one set of data/specifications are used

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

There are minimal issues with missing data because the OASIS submission system rejects assessments with missing values. The provider must then resubmit the assessment.

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*)

There are minimal issues with missing data because the OASIS submission system rejects assessments with missing values. The provider must then resubmit the assessment.

2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias? (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

There are minimal issues with missing data because the OASIS submission system rejects assessments with missing values. The provider must then resubmit the assessment.

3. Feasibility

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

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

3b.2. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

OASIS data collection and transmission is a requirement for the Medicare Home Health Conditions of Participation. Information on bathing status used to calculate this measure is recorded in the relevant OASIS items embedded in the agency's clinical assessment as part of normal clinical practice. OASIS data are collected by the home health agency during the care episode and transmitted electronically to the CMS national OASIS repository. No issues regarding availability of data, missing data, timing or frequency

of data collection, patient confidentiality, time or cost of data collection, feasibility or implementation have become apparent since OASIS-C was implemented 1/1/2010.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

Not Applicable

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting Home Health Compare http://www.cms.gov/HomeHealthCompare/search.asp Home Health Compare http://www.cms.gov/HomeHealthCompare/search.asp Quality Improvement (Internal to the specific organization) Home Health Star Ratings https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits/HHQIHomeHealthStarRatings.html

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

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

The Home Health Compare website is federal government website managed by the Centers for Medicare & Medicaid Services (CMS). It provides information to consumers about the quality of care provided by Medicare-certified home health agencies throughout the nation. The measures reported on Home Health Compare includes all Medicare-certified agencies with at least 20 home health quality episodes. In the 12-month period ending December 31, 2016, there were 9,146 such agencies (81.5 percent of the 11,221 agencies with at least one quality episode) that met the measure denominator criteria for reporting of Improvement in Bathing. This included 4,519,611 episodes of care nationally.

CMS's Home Health Quality Initiative "Outcome Quality Measure Report" provides all Medicare-certified home health agencies with opportunities to use outcome measures for outcome-based quality improvement. The report allows agencies to benchmark their performance against other agencies across the state and nationally, as well as their own performance from prior time periods. All Medicare-certified home health agencies can access their Outcome Quality Measure Reports via CMS's online CASPER system.

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

Not Applicable

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

Not Applicable

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

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

All home health agencies with at least 20 qualifying episodes receive quarterly measure reports on all of their publicly-reported measures. In addition, providers can run on-demand, confidential reports showing individual measure results and national averages, through CMS' CASPER system. There is an email box that HHAs may submit questions to as well as a website on which the latest measure updates are posted.

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

All home health agencies with at least 20 qualifying episodes receive quarterly measure reports on all of their publicly-reported measures. In addition, providers can run on-demand, confidential reports showing individual measure results and national averages, through CMS' CASPER system. There is an email box that HHAs may submit questions to as well as a website on which the latest measure updates are posted. The OASIS Guidance Manual describes the OASIS-based reports that are available as well as the sources of information for the reports. Instructions on using the reports for quality monitoring are provided, illustrated with sample reports from a hypothetical home care agency. It is designed to help home health agencies make use of the reports for monitoring and improving quality of care. Additionally, home health quality reporting program training was held in 2017.

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

Describe how feedback was obtained.

Home health agencies receive quarterly measure reports on all of their measures. There is an email box that HHAs may submit questions to as well as a website on which the latest measure updates are posted. Because of the changes made to the OASIS in OASIS D (effectively January 1, 2019), risk models for publically reported outcome measures have been updated. CMS will makes available information about risk models and covariates on the website and the updated models will be available soon.

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

There is an email box that HHAs may submit regarding quality measures; all questions and responses are captured in an Access database for analysis and CMS receives quarterly reports on questions submitted. Thematic issues arising from the mailbox inform guidance to providers. As in 4a2.2.1.

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

There haven't been any requests for measure modification, nor any modifications made.

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

Not applicable for this time period.

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

Tables 4 and 5 in the "Importance to Report" attachment show observed and predicted measure performance for population groups, respectively. For all population groups, measure performance has improved over time. The greatest improvement in measure performance between 2013 and 2016 for each population subgroup was for:

- Females (66.9 percent in 2013 to 74.0 percent in 2016)
- Whites (67.1 percent in 2013 to 74.8 percent in 2016)
- Under 65 (62.5 percent in 2013 to 69.5 percent in 2016)
- Disabled (63.2 percent in 2013 to 70.3 percent in 2016 – similar to not disabled)
- Not dual (68.0 percent in 2013 to 75.2 percent in 2016 – similar to dual)
- Large HHAs (53.2 percent in 2013 to 63.4 percent in 2016)
- HHAs in the South (66.2 percent in 2013 to 74.5 percent in 2016)

The subgroup with the smallest improvement in performance during this time period was for patients served by small HHAs (bottom 25th percentile in size). Performance for this subgroup only improved from 55.0 percent in 2013 to 56.2 percent in 2016. Note that the number of episodes for small HHAs was only 31,332 during 2016 (or 0.81 percent of episodes for which this measure is available).

There was generally fairly large improvement in measure performance during the 2013 to 2016 period. Overall, improvement was 6.9 percentage points and most population subgroups saw this level of improvement. The largest improvement occurred from 2015 to 2016 – more than half (4.4 percentage points) of the 2013-2016 improvement occurred between 2015 and 2016. We expect to see a similar phenomenon between 2016 and later years. This is likely due to the introduction of several initiatives that incorporate this measure – the Quality of Patient Care (QoPC) Star Ratings, a composite of this

measure and several others that has been publicly reported on Home Health Compare since July 2015 and Home Health Value Based Purchasing (HHVBP). HHVBP began in 2016 and involves nine states. Several participating states encompass a large number of HHAs and providers in other states may be anticipating the expansion of this model.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

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

Recent improvement in this measure has been relatively large compared to historical trends. We believe these large improvements are due to the implementation of two initiatives that involve this measure – the QoPC Star Ratings and HHVBP – beginning in 2015 and 2016.

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

We do not report any unexpected benefits from implementation of this measure at this time.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

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

0430 : Change in Daily Activity Function as Measured by the AM-PAC:

2613 : CARE: Improvement in Self Care

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):
Are the measure specifications harmonized to the extent possible?

No

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

see 5b.1.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

A search using the NQF QPS indicated there are no other endorsed measures that report on rates of improvement in bathing in the home health population. Change in Daily Activity Function as Measured by the AM-PAC (NQF #0430) is a measure of reported changes in patient functioning in the areas of feeding, meal preparation, hygiene, grooming, and dressing as measured by the Activity Measure for Post-Acute Care (AM-PAC), a functional status assessment instrument developed specifically for use in facility and community dwelling post-acute care (PAC) patients. However, the AM-PAC measure is focused on overall functioning (not just bathing), and is calculated using data that are not currently collected in the home health setting.

CARE: Improvement in Self Care (NQF# 2613) is a measure of self-care based on the subscale of the Continuity Assessment and Record Evaluation (CARE) Tool and information from the admission MDS 3.0 assessment. The measure specifications and exclusions don't currently apply to home health.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** 0174_Bath_Importance_to_Report_Tables.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services

Co.2 Point of Contact: Joan, Proctor, Joan.Proctor2@cms.hhs.gov, 443-526-6938-

Co.3 Measure Developer if different from Measure Steward: Centers for Medicare & Medicaid Services

Co.4 Point of Contact: Joan, Proctor, Joan.Proctor2@cms.hhs.gov, 443-526-6938-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Not Applicable

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2004

Ad.3 Month and Year of most recent revision: 11, 2011

Ad.4 What is your frequency for review/update of this measure? Annual

Ad.5 When is the next scheduled review/update for this measure? 09, 2015

Ad.6 Copyright statement: NA

Ad.7 Disclaimers: NA

Ad.8 Additional Information/Comments: NA