**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b6)**

**Measure Title**: Percent of Residents with a Urinary Tract Infection (long stay)

**Date of Submission**: 1/17/2014

**Type of Measure:**

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

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| ☐ administrative claims | ☐ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| ☐ abstracted from electronic health record | ☐ abstracted from electronic health record |
| ☐ eMeasure (HQMF) implemented in EHRs | ☐ eMeasure (HQMF) implemented in EHRs |
| X other: nursing home Minimum Data Set (MDS) | X other: nursing home Minimum Data Set (MDS) |

**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).

Nursing Home MDS 2.0 and 3.0

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

Q3 of 2003 to Q3 of 2013

**1.4. What levels of analysis were tested?** *(testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan)*

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| ☐ individual clinician | ☐ individual clinician |
| ☐ group/practice | ☐ group/practice |
| X hospital/facility/agency | ☐ hospital/facility/agency |
| ☐ health plan | ☐ health plan |
| ☐ other: Click here to describe | ☐ other: Click here to describe |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)?** (identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample)

There were multiple studies involved in testing and analysis (described in greater detail below):

1. University of Colorado analysis: 10% random sample of all Medicare certified nursing facilities (n = 1,603)
2. RAD comparison of MDS 2.0 and MDS 3.0: 71 community nursing homes in 8 states, 19 VA nursing homes
3. RTI Analysis: The measure is not an estimate based on samples; rather it includes all nursing home residents nationwide who meet the inclusion criteria. All nursing facilities that had sufficient sample size (n ≥ 30) to report on this measure (N = 13,640 in Q2 2013)

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)?** *(identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample)*

1. University of Colorado analysis: n not reported for reliability analysis
2. RAND comparison of MDS 2.0 and MDS 3.0: 1,402 nursing home residents
3. RTI Analysis: All residents that meet denominator inclusion criteria (1,108,999 in Q2 2013)

**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.**

## **Data Sample for Reliability Testing**

Three sets of tests of the reliability of the urinary tract infection measure have been conducted.

First, the University of Colorado used national facility-level quality measure data from Quarter 3 (Q3) 2003 through Q3 2006 that came from the Quality Improvement and Evaluation System (QIES) MDS Express Reports on the CMS intranet; Online Survey, Certification, and Reporting (OSCAR) data related to facility characteristics (e.g., state, resident census, number of beds, staffing) and certification survey results from QIES Workbench.(1) A 10% random sample of all Medicare-certified nursing facilities was also downloaded from MDS assessment records. Analyses were based on complete MDS data from January 2005 through March 2006, as well as nearly complete data for April 2006 and partial data for May and June 2006.

Second, testing of the reliability of MDS 3.0 data items underlying the urinary tract infection quality measure as well as a comparison with the MDS 2.0 quality measures were conducted by RAND as part of the MDS 3.0 development process.(2) A representative sample of for-profit and not-for-profit facilities and hospital-based and free-standing facilities was recruited for the study, which included 71 community nursing homes in 8 states, 19 VA nursing homes, and 1,402 nursing home residents for the urinary tract infection quality measure.

Third, RTI has continued to analyze reliability and validity of the measures as part of ongoing maintenance. These analyses include all nursing home residents that are eligible for denominator inclusion for the measure in all facilities that have adequate sample size (at least 30 residents) to report.

## **Data Sample for Validity Testing**

Three studies have examined the validity of the urinary tract infection measure.

First, the University of Colorado used national facility-level quality measure data from Quarter 3 (Q3) 2003 through Q3 2006 that came from the Quality Improvement and Evaluation System (QIES) MDS Express Reports on the CMS intranet; Online Survey, Certification, and Reporting (OSCAR) data related to facility characteristics (e.g., state, resident census, number of beds, staffing) and certification survey results from QIES Workbench.(1) A 10% random sample of all Medicare-certified nursing facilities was also downloaded from MDS assessment records. Analyses were based on complete MDS data from January 2005 through March 2006, as well as nearly complete data for April 2006 and partial data for May and June 2006.

Second, testing of the reliability of MDS 3.0 data items underlying the urinary tract infection quality measure as well as a comparison with the MDS 2.0 quality measures were conducted by RAND as part of the MDS 3.0 development process.(3) A representative sample of for-profit and not-for-profit facilities and hospital-based and free-standing facilities was recruited for the study, which included 71 community nursing homes in 8 states, 19 VA nursing homes, and 1,402 nursing home residents for the urinary tract infection quality measure.

Third, RTI has continued to analyze reliability and validity of the measures as part of ongoing maintenance. These analyses include all nursing home residents that are eligible for denominator inclusion for the measure in all facilities that have adequate sample size (at least 30 residents) to report.

1. Brega A, Hittle D, Goodrich G, Kramer A, Conway K, Levy C. Empirical review of publicly reported nursing home quality measures. Denver: Division of Health Care Policy and Research University of Colorado at Denver; Abt Associates, Inc, 2007.
2. Saliba D, Buchanan J. Development and Validation of a Revised Nursing Home Assessment Tool: MDS 3.0. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation, Apr 2008. Available from http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf.
3. Stevenson KB, Moore JW, Sleeper B. Validity of the Minimum Data Set in identifying urinary tract infections in residents of long-term care facilities. J Am Geriatr Soc. 2007;52: 707-711.

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

# 2a2. RELIABILITY TESTING

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)

X **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)   
X **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*)

**Data Element Reliability**

The national test of MDS 3.0 items examined agreement between assessors (reliability).(1) Quality Improvement Organizations (QIOs) were employed to identify gold-standard (research) nurses and recruit community nursing homes to participate in the national evaluation. The gold-standard nurses were trained in the MDS 3.0 instrument and, in turn, they trained a facility nurse from each participating nursing home in their home states. Residents participating in the test were selected to capture a representative sample of short and long-stay residents. Quality measures using the MDS 2.0 and the MDS 3.0 were calculated and then compared, with correlations and Kappas calculated.

**Performance Measure Score Reliability**

Two sets of analytic methods were used for performance measure score reliability. First, the University of Colorado examined the percentage of facilities that had a change in ranking from one quarter to the next of at least three deciles.(2) This indicator of stability was computed for each of the twelve pairs of adjacent quarters for which data were available (2003 Q3 through 2006 Q3).

In a similar analysis, RTI examined the extent to which relative facility rank changed quarterly on this quality measure between Quarter 4 of 2011 and Quarter 1 of 2013. Dramatic changes in facility QM scores from one quarter to the next are more likely an indicator of measure instability than of great improvement or decline in facility performance.

1. Saliba D, Buchanan J. Development and Validation of a Revised Nursing Home Assessment Tool: MDS 3.0. Contract No. 500-00-0027/Task Order #2. Santa Monica, CA: Rand Corporation, Apr 2008. Available from http://www.cms.hhs.gov/NursingHomeQualityInits/Downloads/MDS30FinalReport.pdf.
2. Brega A, Hittle D, Goodrich G, Kramer A, Conway K, Levy C. Empirical review of publicly reported nursing home quality measures. Denver: Division of Health Care Policy and Research University of Colorado at Denver; Abt Associates, Inc, 2007.

**2a2.3. For each level checked above, what were the statistical results from reliability testing**? (e.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis)

**Data Element Reliability**

In their testing of the MDS 3.0, RAND compared the results on the nursing home quality measures using the MDS 3.0 and the MDS 2.0, both at the individual resident level and at the facility level.(2) At the resident level, the urinary tract infection rate using the MDS 2.0 was 10.0% and using the MDS 3.0 was 7.5%; the Kappa was 0.70 and the correlation was 0.71. Kappa is a statistical measure of inter-rater agreement for qualitative data, ranging from 0.0 to 1.0. A rating of 0.70 is considered “substantial agreement.” At the facility level, the MDS 2.0 rate of urinary tract infections was 10.2% and the MDS 3.0 rate was 7.3%, with a correlation of 0.80, which is quite high.

**Performance Measure Score Reliability**

In terms of measure stability, the University of Colorado examined the percentage of facilities that had a change in ranking of at least three deciles from one quarter to the next.(2) For urinary tract infection, 30.4% of facilities had a three-decile-or-more change from one quarter to the next quarter. The range of stability measures across the 12 comparisons was very small (i.e., the difference between the maximum and minimum values), indicating that measure stability is quite constant over time. For urinary tract infections, the minimum percentage was 29.9%, and the maximum percentage was 31.0%.

Results of the RTI analysis are presented in **Figure 1**, whichillustrates quarterly change in facility rank in 2012. For each pair of quarters, more than half of facilities maintain the same rank (within one decile) from one quarter to the next. Changes of one to two deciles are fairly common, occurring among approximately twenty percent of facilities in each pair of quarters. The most recent transition over the two most recent quarters was the least stable of the quarters examined, with 15.3 percent of facilities seeing rank changes of more than three deciles.

Figure 1   
Distribution of change in facility rank from one quarter to the next, QM #0684 Percent of Residents with a Urinary Tract Infection

Analysis date: 8/5/2013

NOTES: Denominators for these proportions reflect facilities that meet minimum requirements for public reporting this QM in both Quarters.

SOURCE: MDS 3.0, RTI analysis of MDS 3.0 episode files for Quarter 4, 2011 to Quarter 1, 2013 2013(\quarter\_9\_10\db241\db241\_request.log)

1. Stevenson KB, Moore JW, Sleeper B. Validity of the Minimum Data Set in identifying urinary tract infections in residents of long-term care facilities. J Am Geriatr Soc. 2007;52: 707-711.
2. Brega A, Hittle D, Goodrich G, Kramer A, Conway K, Levy C. Empirical review of publicly reported nursing home quality measures. Denver: Division of Health Care Policy and Research University of Colorado at Denver; Abt Associates, Inc, 2007.

**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?)

Most studies have shown a high level of reliability for this measure. The most recent examination by RTI, however, indicates that facility rank may not be as reliable as earlier studies had shown. Although these recent results indicate a degree of instability, the change may be explained by relatively low prevalence and narrow interquartile range that makes relatively small changes in score appear to be large changes in rank.

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

# 2b2. VALIDITY TESTING

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)

X **Critical data elements** (*data element validity must address ALL critical data elements*)

X **Performance measure score**

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

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

**Critical Data Elements**

In a study of the validity of the urinary tract infection quality measure, Stevenson, Moore and Sleeper recruited 16 Idaho nursing homes to voluntary participate in a CMS-funded performance improvement project to reduce inappropriate antimicrobial prescribing for urinary tract infections from July 2001 to June 2002.(1)

**Performance Measure Score**

Two studies examined the validity of the urinary tract infection measure.

First, the University of Colorado used national facility-level quality measure data from Quarter 3 (Q3) 2003 through Q3 2006 that came from the Quality Improvement and Evaluation System (QIES) MDS Express Reports on the CMS intranet; Online Survey, Certification, and Reporting (OSCAR) data related to facility characteristics (e.g., state, resident census, number of beds, staffing) and certification survey results from QIES Workbench.(2) A 10% random sample of all Medicare-certified nursing facilities was also downloaded from MDS assessment records. Analyses were based on complete MDS data from January 2005 through March 2006, as well as nearly complete data for April 2006 and partial data for May and June 2006.

Second, RTI examined several aspects of validity: Whether facilities’ scores on this measure correlated with scores on related measures; whether missing data on resident assessments have an effect on facilities’ measure scores; and whether there is any evidence that seasonal variation is a threat to validity.

Systematic assessment of face validity: In a clinical review of the nursing home QMs using the MDS 2.0, a Technical Expert Panel (TEP) concluded that the urinary tract infection quality measure is a “valuable source of information for nursing homes”(2) ([2](#_ENREF_17)). The measure prompts facilities to examine their approach to perineal care and their general infection rate. These infections have the potential for significant morbidity and mortality (3). Infections increase the use of medical care and costs. Because many UTIs are related to catheters, this quality measure provides an additional incentive for the facility to monitor its catheter use (4) ([Gould et al., 2009](#_ENREF_64)).

1. Stevenson KB, Moore JW, Sleeper B. Validity of the Minimum Data Set in identifying urinary tract infections in residents of long-term care facilities. J Am Geriatr Soc. 2007;52: 707-711.
2. Brega A, Hittle D, Goodrich G, Kramer A, Conway K, Levy C. Empirical review of publicly reported nursing home quality measures. Denver: Division of Health Care Policy and Research University of Colorado at Denver; Abt Associates, Inc, 2007.
3. Nicolle LE, SHEA Long-Term Care Committee. Shea Position Paper: Urinary Tract Infections in Long-Term Care Facilities. Infect Control Hosp Epidemiol. 2001;22:167-175. Available: http://www.shea-online.org/assets/files/other\_papers/utis\_in\_ltcf\_2001.pdf
4. Gould CV, Umscheid CA, Agarwal R, Kuntz G, Pegues DA; the Healthcare Infection Control Practices Advisory Committee. Guideline for prevention of catheter-associated urinary tract infections 2009. Atlanta: Centers for Disease Control and Prevention, 2009. Available from http://www.cdc.gov/ncidod/dhqp/pdf/guidelines/CAUTI\_Guideline2009final.pdf.

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

**Critical Data Elements**

In a study of 16 nursing homes in Idaho, researchers examined the validity of the MDS 2.0 urinary tract infection quality measure by comparing the MDS results to those of an active, prospective surveillance program in the facilities.(1) While almost all of the urinary tract infections were identified by the MDS, the measure also identified a substantial number of false positives. The estimated sensitivity (the proportion of residents listed with a urinary tract infection with an actual urinary tract infection ) of the MDS was 57.9% and specificity (proportion of residents not listed as having a urinary tract infection that did not have a urinary tract infection) was 86.5%. Given the importance of infection control in nursing homes, it is preferable to cast a fairly wide net to identify all persons who need treatment and to motivate facilities to improve their practices.

**Performance Measure Score**

Correlations. In an analysis by the University of Colorado, the urinary tract infection measure had correlations of 0.12 or less with other publicly reported nursing home quality measures.(1) The only correlation that was higher was 0.28 with indwelling catheter, which would be expected given that both measures involve the urinary tract.

Facilities should perform similarly on quality measures that reflect the quality of similar care processes. RTI examined whether a nursing home’s percentile rank on related quality measures was correlated, restricting analyses to facilities that meet public reporting sample size requirements and using risk adjusted scores where applicable.

* Related measures examined: QM # 0685 Percent of Low Risk Residents Who Lose Control of their Bowels or Bladder (Long Stay) and QM #0686 Percent of Residents Who Have/Had a Catheter Inserted and Left in Their Bladder.

Results: RTI found a weak but statistically significant correlation between the facility percentile ranking of this measure and both QM #0685 (r = 0.111, p < 0001) and QM #0686 (r = 0.100, p < 0001).(2)

Missing Data and Exclusions. Missing data represent a potential threat to the validity of a quality measure. Bias may be introduced if missing data are systematically associated with resident or facility characteristics.

* In Quarter 2 2013, 22,759 long-stay residents (1.9%) were excluded from the calculation of this measure. Of these, 838 (0.1 percent of long-stay residents) were excluded because of missing data on items used to calculate the measure, and 21,931 (1.9% of long-stay residents) were excluded because the target assessment was an admission, 5-day, or readmission/return assessment (**Table 1**). Facilities would be penalized unfairly if individuals who developed UTIs prior to arrival in the nursing facility (i.e., prior to admission, 5-day, or readmission) were included in the measure.

Table 1  
Target assessments used in the calculation of QM #0684 Percent of Residents with a Urinary Tract Infection (Long Stay)

| Resident episodes in the reporting period | Frequencya | Percentagea |
| --- | --- | --- |
| Included | 1,154,742 | 98.0% |
| Excluded—Missing data | 838 | 0.1% |
| Excluded—Wrong target assessment (admission, 5-day, or readmission/return assessment) | 21,931 | 1.9% |
| Total number of long-stay resident episodes | 1,177,501 | — |

NOTES:

a Column values may not add up to total since a resident episode can meet more than one exclusion criteria. Percentage column reflects percentage of target assessments in each category out of total resident episodes in long-stay population.

Analysis date: 8/27/13

SOURCE: RTI analysis of MDS 3.0 episode file for Quarter 1, 2013 ([\quarter\_10\_11\db267\_request.log](file:///\\\\rtimas04\\hser\\Project\\0211942\\001%20%20MIDS-NHQ\\004%20Testing%20and%20Validation\\common\\ykaganova\\db\\quarter_4_5\\db127_request_v1.log)).

**Table 2** provides summary statistics for an analysis of the distribution of missing data rates for facilities reporting on this measure. Missing data are rare for this measure: the facility-level mean missing data rate for items used to calculate this measure is 0.1 percent; over 90 percent of facilities have no missing data.

Table 2  
Distribution of facility-level missing rate for QM #0684 Percent of Residents with a Urinary Tract Infection (Long Stay)

| *n* | Mean | Std dev. | 10th  percentile | 25th  percentile | 50th  percentile | 75th  percentile | 90th  percentile |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 13,773 | 0.1% | 0.8% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |

NOTES:

*n* = number of facilities who have data for the numerator and denominator of this QM before assessment exclusion criteria are applied; facilities are included regardless of whether they meet the minimum sample size for reporting.

Analysis date: 8/6/2013. SOURCE: RTI analysis of MDS 3.0 episode file for Quarter 1, 2013 (\quarter\_10\_11\db243\_request.log)

* We also examined the relationship between missing data and QM scores.
  + **Table 3** shows the distributions of facility-level missing data rates stratified by quartile of QM scores for this measure. The mean facility-level missing rate is less than or equal to 0.1 percent for each quartile of QM score.
  + There is a significant but weak correlation between missing data and QM scores for this measure (r = .031, p < .001).
  + Facilities with higher levels of missing data appear to have slightly greater proportions of residents with urinary tract infections, but the overall missing data rate is likely too small to have any meaningful implications for this measure.

Table 3  
Distribution of facility-level missing rate for QM #0684 Percent of Residents with a Urinary Tract Infection (Long Stay)

| Quartile of QM score | *n* | Mean | Std dev. | 10th  percentile | 25th  percentile | 50th  percentile | 75th  percentile | 90th  percentile |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0%–25% | 3,445 | 0.0% | 0.4% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| 26%–50% | 3,442 | 0.0% | 0.4% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| 51%–75% | 3,448 | 0.1% | 1.3% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| 76%–100% | 3,438 | 0.1% | 0.8% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |

NOTES:

Total *n* (13,773) = number of facilities who have data for the numerator and denominator of this QM; facilities are included regardless of whether they meet the minimum sample size for reporting.

Analysis date: 8/6/2013

SOURCE: RTI analysis of MDS 3.0 episode file for Quarter 1, 2013.

Seasonal variation. Another potential threat to the validity of a QM is seasonal variation. If a QM score varies substantially from quarter to quarter in a consistent temporal pattern that corresponds corresponding to changes in seasons, it is possible that the validity of the measure is compromised due to influences beyond nursing home control. Since seasonal variation might play a role in determining facility scores, we used a six month average, examining the national mean and median for this QM score between Quarter 1 2011 and Quarter 1 2013.

* The results are presented in **Figure 2**. The national facility-level mean peaked in Q3 2011 at 7.9 percent, as did the national facility-level median (6.9 percent). Since then, both the mean and the median have declined steadily, with the national mean now at 6.5 percent and the median at 5.6 percent, indicating that there is no seasonal variation.

Figure 2  
Seasonal (quarterly) variation in QM #0684 Percent of Residents with a Urinary Tract Infection (long stay)

SOURCE: RTI analysis of MDS 3.0 episode file for Quarter 1, 2011 – Quarter 1, 2013 \qm\_quarter\_1\_2\complete\nh\_018\_10.log, \qm\_quarter\_2\_3\complete\nh\_018\_10.log \qm\_quarter\_3\_4\complete\nh\_018\_10.log, \qm\_quarter\_4\_5\complete\nh\_018\_10.log, \qm\_quarter\_5\_6\complete\nh\_018\_10.log, \qm\_quarter\_6\_7\complete\nh\_018\_10.log, \qm\_quarter\_8\_9\complete\nh\_018\_10.log, \qm\_quarter\_9\_10\complete\nh\_018\_10.log,

1. Stevenson KB, Moore JW, Sleeper B. Validity of the Minimum Data Set in identifying urinary tract infections in residents of long-term care facilities. J Am Geriatr Soc. 2007;52: 707-711.
2. SOURCE: RTI analysis of MDS 3.0 episode file for Quarter 1, 2013 (\quarter\_10\_11\db263\_request\db263\_request.log)
3. Brega A, Hittle D, Goodrich G, Kramer A, Conway K, Levy C. Empirical review of publicly reported nursing home quality measures. Denver: Division of Health Care Policy and Research University of Colorado at Denver; Abt Associates, Inc., 2007.

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

1. This measure is significantly correlated with each of the other measures related to gastrourinary health.
2. Missing data do not present a threat to the validity of this measure
3. There is no apparent threat of seasonal variation to the validity of this measure.

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

# 2b3. EXCLUSIONS ANALYSIS

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

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

Exclusion rates were examined by reason for exclusion, (missing data and admission assessments) were calculated. Facility scores were both stratified by and correlated with missing data rates at the facility level.

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

* In Quarter 1 2013, 22,759 long-stay residents (1.9%) were excluded from the calculation of this measure. Of these, 838 (0.1 percent of long-stay residents) were excluded because of missing data on items used to calculate the measure, and 21,931 (1.9% of long-stay residents) were excluded because the target assessment was an admission, 5-day, or readmission/return assessment (Table 1, above). Facilities would be penalized unfairly if individuals who developed UTIs prior to arrival in the nursing facility (i.e., prior to admission, 5-day, or readmission) were included in the measure.
* Table 2 (above) shows the distributions of facility-level missing data rates stratified by quartile of QM scores for this measure. The mean facility-level missing rate is less than or equal to 0.1 percent for each quartile of QM score.
* There is a significant but weak correlation between missing data and QM scores for this measure (r = .031, p < .001).
* Facilities with higher levels of missing data appear to have slightly greater proportions of residents with urinary tract infections, but the overall missing data rate is likely too small to have any meaningful implications for this measure.

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

There are very few missing data on items used to calculate this measure, and thus there are very few exclusions. Therefore, exclusions are unlikely to impact performance results.

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

# 2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

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

X **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of risk factors **risk factors**

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

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

**2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.**

The measure is not risk adjusted through a statistical model. However, the measure only applies to long-stay residents. The measure is limited to the long stay population because post-acute care patients may have developed urinary tract infection in the hospital rather than the nursing facility. Urinary tract infections have relatively high prevalence across the continuum of care and there are no obvious conditions for which risk adjustment is appropriate. In particular, urinary tract infections are often associated with catheter use, which is often inappropriate.(1)

Thus, risk adjusting for the proportion of residents who have catheters would not be desirable.

1. Gould CV, Umscheid CA, Agarwal R, Kuntz G, Pegues DA; the Healthcare Infection Control Practices Advisory Committee. Guideline for prevention of catheter-associated urinary tract infections 2009. Atlanta: Centers for Disease Control and Prevention, 2009. Available from http://www.cdc.gov/ncidod/dhqp/pdf/guidelines/CAUTI\_Guideline2009final.pdf.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk** (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities)

Not applicable

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

Not applicable

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

Not applicable

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

Not applicable

**if stratified, skip to** [**2b4.9**](#question2b49)

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

Not applicable

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

Not applicable

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

Not applicable

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

Not applicable

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

Not applicable

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

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

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The measure is not an estimate based on samples; rather it includes all nursing home residents nationwide who meet the inclusion criteria.

However, the method for determining meaningful differences in facility performance has not yet been determined. For several publicly reported measures, CMS currently estimates an interval estimate for each provider rate to characterize the amount of uncertainty associated with each rate and compares the interval estimate to the national mean for the outcome.

Following prior CMS work, in this analysis we characterize the proportion of facilities that had scores that were significantly lower than the national mean, significantly higher than the national mean, or not significantly different from the mean in order to see how well the measure distinguishes high-performing from low-performing facilities. This was done by finding the 95 percent confidence interval for each facility and observing where the national mean fell in relation to those intervals. Please note that it is not possible to calculate confidence intervals for facilities with scores of 0 percent or 100 percent (because the standard error of either proportion is always 0). These facilities were treated as having intervals of either [0,0] or [1,1].

This analysis was also stratified by facility size (using deciles of number of residents in the denominator as a measure of volume).

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The distribution, for the second quarter of 2013, of QM #0684 Percent of Residents with a Urinary Tract Infection (Long Stay) is shown in **Table 4** below. The measure showed a range of variability across facilities nationally (n = 13,640). The distribution of facility level scores had a slight, positive skew, with a mean of 6.2 percent and a median of 5.3 percent. The interquartile range is 6.4 percent. At least ten percent of facilities had perfect scores of 0 percent for this measure, suggesting the possibility of a floor effect for this measure. This analysis was restricted to facilities with at least 30 residents in the denominator, the minimum requirement for public reporting.

Table 4  
Percentile distribution of scores for QM #0684 Percent of Residents   
with a Urinary Tract Infection (Long Stay)

|  |  |
| --- | --- |
| Percentile | Value |
| N (facilities) | 13,640 |
| k (residents) | 1,108,999 |
| mean | 6.2% |
| SD | 5.0% |
| min | 0.0% |
| max | 38.5% |
| Interquartile Range | 6.4% |
| 10th % | 0.0% |
| 20th % | 2.0% |
| 30th % | 3.0% |
| 40th % | 4.2% |
| 50th % | 5.3% |
| 60th % | 6.5% |
| 70th % | 8.1% |
| 80th % | 9.9% |
| 90th % | 12.9% |

NOTES:

n = number of facilities that meet minimum requirements for public reporting this quality measure.

QM scores are reported at the facility level.

Analysis date: 1/14/2014

SOURCE: RTI analysis of MDS 3.0 episode file for Quarter 2, 2013 (quarter\_10\_11\db277\db277\_request.log).

To identify meaningful differences in performance between providers, we estimated 95 percent confidence intervals around the providers’ QM scores allowing for comparison with the national average. These results are summarized in **Table 5** below. We found that 32 percent of nursing homes overall were significantly different than the national average rate for QM# 0684. The percent of nursing homes which were significantly different increases as facility size increases; for example, 23 percent of nursing homes in the smallest decile based on volume were significantly different compared to 42.9 percent significantly different in decile 10, the largest nursing homes.

Table 5   
Proportion Nursing Homes Statistically Significantly Different from National Mean, Overall and by Deciles of Facility Denominator Count

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Volume Deciles  (min-max) | Number of facilities | Number significantly different | Percent significantly different | Percent significantly higher (worse) | Percent significantly lower (better) |
| Decile 1  (30-39) | 1,406 | 324 | 23.0% | 5.4% | 17.6% |
| Decile 2  (40-47) | 1,392 | 307 | 22.1% | 5.3% | 16.7% |
| Decile 3  (48-55) | 1,405 | 464 | 33.0% | 4.5% | 28.5% |
| Decile 4  (56-63) | 1,297 | 411 | 31.7% | 5.2% | 26.5% |
| Decile 5  (64-72) | 1,456 | 387 | 26.6% | 6.5% | 20.1% |
| Decile 6  (73-80) | 1,231 | 433 | 35.2% | 7.2% | 27.9% |
| Decile 7  (81-90) | 1,383 | 498 | 36.0% | 8.1% | 27.9% |
| Decile 8  (91-104) | 1,389 | 470 | 33.8% | 7.3% | 26.5% |
| Decile 9  (105-131) | 1,342 | 542 | 40.4% | 8.7% | 31.7% |
| Decile 10  (132-813) | 1,339 | 575 | 42.9% | 6.4% | 36.5% |
| Overall | 13,640 | 4,411 | 32.3% | 6.5% | 25.9% |

Analysis date: 1/9/2014

Analysis is restricted to facilities with minimum denominator size for public reporting (n=30).

SOURCE: RTI analysis of MDS 3.0 episode file for Quarter 1, 2013 (\quarter\_10\_11\db279\db279\_request.log).

The last two columns present the percent that are significantly higher (worse) and significantly lower (better) than average. Across all deciles, the proportion of nursing homes with scores significantly better than the national average is higher, an imbalance presumably driven by facilities with perfect scores and the skewed distribution of facility scores.

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)  
**From**

The proportion of facilities that are significantly better than national average and the proportion of facilities that are significantly worse than national average increases as the number of residents in the facility increases. This may be due to the fact that the standard errors decrease as the number of residents increases, but may also be evidence that reliability is better for larger facilities. Finally, we point out that while this approach for assessing meaning differences among facilities using interval estimates and a national mean is an accepted methodological approach in current quality measure reporting, a ‘gold standard’ method for determining meaningful differences in facility performance has not been established.

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

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

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

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

Not applicable

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

Not applicable

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

Not applicable

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (*i.e., what do the results mean and what are the norms for the test conducted*)

Not applicable