April 2016

Hospice Quality Reporting Program: Executive Summary of Measure Testing and Validation

Prepared for

Center for Clinical Standards and Quality
Centers for Medicare & Medicaid Services
7500 Security Boulevard
Baltimore, MD 21244-1850

CMS Contract No. HHSM-500-2013-13015I

RTI INTERNATIONAL

HOSPICE FACILITY QUALITY REPORTING PROGRAM: EXECUTIVE SUMMARY OF MEASURE TESTING AND VALIDATION

RTI International

CMS Contract No. HHSM-500-2013-13015I

April 2016

This project was funded by the Centers for Medicare & Medicaid Services under contract no. HHSM-500-2013-13015I. The statements contained in this report are solely those of the authors and do not necessarily reflect the views or policies of the Centers for Medicare & Medicaid Services. RTI assumes responsibility for the accuracy and completeness of the information contained in this report.

EXECUTIVE SUMMARY

Section 3004 of the Patient Protection and Affordable Care Act (ACA) authorizes the Secretary of Health and Human Services to establish a quality reporting program for hospices. The ACA specifies that, for fiscal year (FY) 2014 and each subsequent FY, hospice programs shall submit data on quality measures (QMs) to the Secretary. The ACA also describes measure endorsement requirements for any measures specified by the Secretary. The Centers for Medicare & Medicaid Services (CMS) implemented the Hospice Quality Reporting Program (HQRP) in the FY 2012 Hospice Wage Index final rule (76 FR 47302-47352). CMS implemented the Hospice Item Set (HIS), a standardized, patient-level data collection instrument, as part of the HQRP in the FY 2014 Hospice Wage Index final rule (78 FR 48234-48281). Medicare-certified hospices are required to submit an HIS-Admission record and an HIS-Discharge record for each patient admitted on or after July 1, 2014.

The current version of the HIS (V1.00.0) collects patient-level data to calculate seven QMs. Six are endorsed by the National Quality Forum (NQF), and one is a modified NQF-endorsed QM.[1] These measures focus on care processes around hospice admission that are either clinically recommended, required in the hospice Conditions of Participation, or both. The current seven HQRP QMs are as follows:

- NQF #1641, Treatment Preferences.
- Modified NQF #1647, Beliefs/Values Addressed (if desired by the patient).
- NQF #1634, Pain Screening.
- NQF #1637, Pain Assessment.
- NQF #1639, Dyspnea Screening.
- NQF #1638, Dyspnea Treatment.
- NQF #1617, Patients Treated with an Opioid Who Are Given a Bowel Regimen.

To test the scientific soundness of the QMs, RTI International conducted analyses that assess the overall performance of the QMs, using the criteria specified in the guidelines for determining measure reliability and validity, set forth in the CMS Measures Management System Blueprint and by the NQF. These analyses address the following key areas: reportability, distribution and variability, reliability, and validity. We also examined the impact of measure exclusions on the measures' performance and disparities in hospice care.

The analyses in this report primarily used hospice patient stays discharged from October 1, 2014, through September 30, 2015. We conducted these analyses using HIS-Admission and -Discharge records with target dates from October 1, 2014, through September 30, 2015. We summarize the findings below.

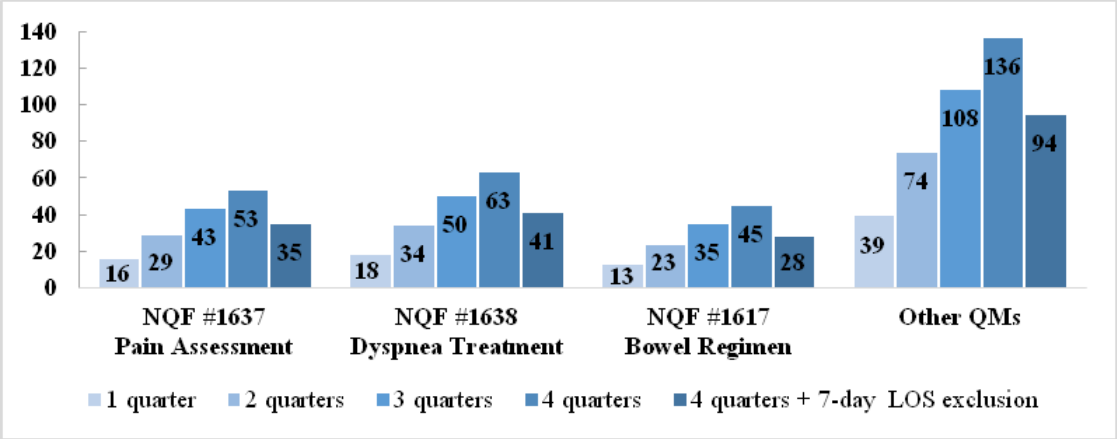
[1] The NQF-endorsed measure specifications indicate that the numerator condition can be met at any time during the patient's hospice episode of care. However, we have opted to include the relevant items for this measure on the HIS-Admission.

Reportability and Length-of-Stay Exclusion

Reportability analysis assesses whether the QM denominator size is large enough to generate statistically reliable scores. We examined the impact of specifications that can affect the denominator size, including the data selection period and the length-of-stay (LOS) exclusion.

- *Data selection period.* We examined the distributions of the QM denominator sizes by four data selection period options: one, two, three, and four quarters of data. The analyses, presented in **Figure ES-1**, show that using four quarters of data produced significantly larger QM denominator sizes among hospices. For example, for four QMs (NQF #1641, Treatment Preferences; NQF #1647, Beliefs/Values; NQF #1634, Pain Screening; and NQF #1639, Dyspnea Screening), the median denominator size was 136 when calculating QM scores with four quarters of data, 39 with one quarter, 74 with two quarters, and 108 with three quarters. The smallest denominator size was for NQF #1617, Bowel Regimen, with a median of 45 using four quarters of data.

Figure ES-1
Median Denominator Size by Data Selection Period and LOS Exclusion



SOURCE: RTI International analysis of HIS data.

NOTE: Four QMs–NQF #1641, Treatment Preferences; NQF #1647, Beliefs/Values; NQF #1634, Pain Screening; and NQF #1639, Dyspnea Screening–have the same denominator size.

- *LOS exclusion*. In the current measure specifications for six of the seven QMs, patient stays shorter than 7 days are excluded from the measure denominator. The 7-day LOS exclusion has been a topic of discussion and concern because a large proportion of hospice stays are less than 7 days. We examined the impact of the LOS exclusion on the measures' reportability and variability by reporting the distributions of the QMs' denominator size and score with and without applying the LOS exclusion. Our findings show that denominator size for the QMs was significantly affected by the current LOS exclusion. For all QMs, the median number of stays per hospice when applying the LOS exclusion was at most about 70 percent of that when not applying the exclusion ***(Figure ES-1)***. Excluding stays that are less than 7 days also excluded some hospices from the QM score calculations because these hospices did not have any qualifying stays.

Our analyses show minimal differences in the distributions of QM scores with and without applying the LOS exclusion. Additionally, clinical practice, confirmed by our analyses, reveals that the care processes targeted by the QMs are performed on the day of admission for the vast majority of patient stays. This suggests that good clinical practice at hospice admission includes the performance of these care processes, even for patients with stays shorter than 7 days. Thus, including these patient stays in QM calculations (i.e., removing the 7-day LOS exclusion) may be appropriate and would not create a burden on hospices.

We recommend using four quarters of data for measure calculation and modifying the current measure specifications by removing the 7-day LOS exclusion to increase denominator size and capture a larger number of hospice stays;[2] because of this recommendation, subsequent analyses on this measure were conducted with no LOS exclusion.

- *Public reporting.* As previously stated, small denominators may not yield statistically meaningful QM scores. Thus, for quality reporting programs, CMS commonly sets a minimum denominator threshold for public reporting. A commonly used minimum denominator size in other quality reporting programs is 20. We examined the impact of this threshold on the number of hospices that would qualify for the public reporting of each QM (***Table ES-1***). We found that when applying a minimum denominator size of 20 stays, using rolling 12 months of data to create QMs would suppress about 10 percent of hospices from public reporting on four out of the seven QMs, and about 20–30 percent from public reporting on the other three QMs.

Table ES-1
Number of Hospices Able to Publicly Report with the Minimum Threshold Applied

QMs	N0	N1	%
NQF #1641, Treatment Preferences	3,922	405	10.33
NQF #1647, Beliefs/Values	3,922	405	10.33
NQF #1634, Pain Screening	3,922	405	10.33
NQF #1637, Pain Assessment	3,883	883	22.74
NQF #1639, Dyspnea Screening	3,922	405	10.33
NQF #1638, Dyspnea Treatment	3,883	786	20.24
NQF #1617, Bowel Regimen	3,843	1,111	28.91

SOURCE: RTI International analysis of HIS data.

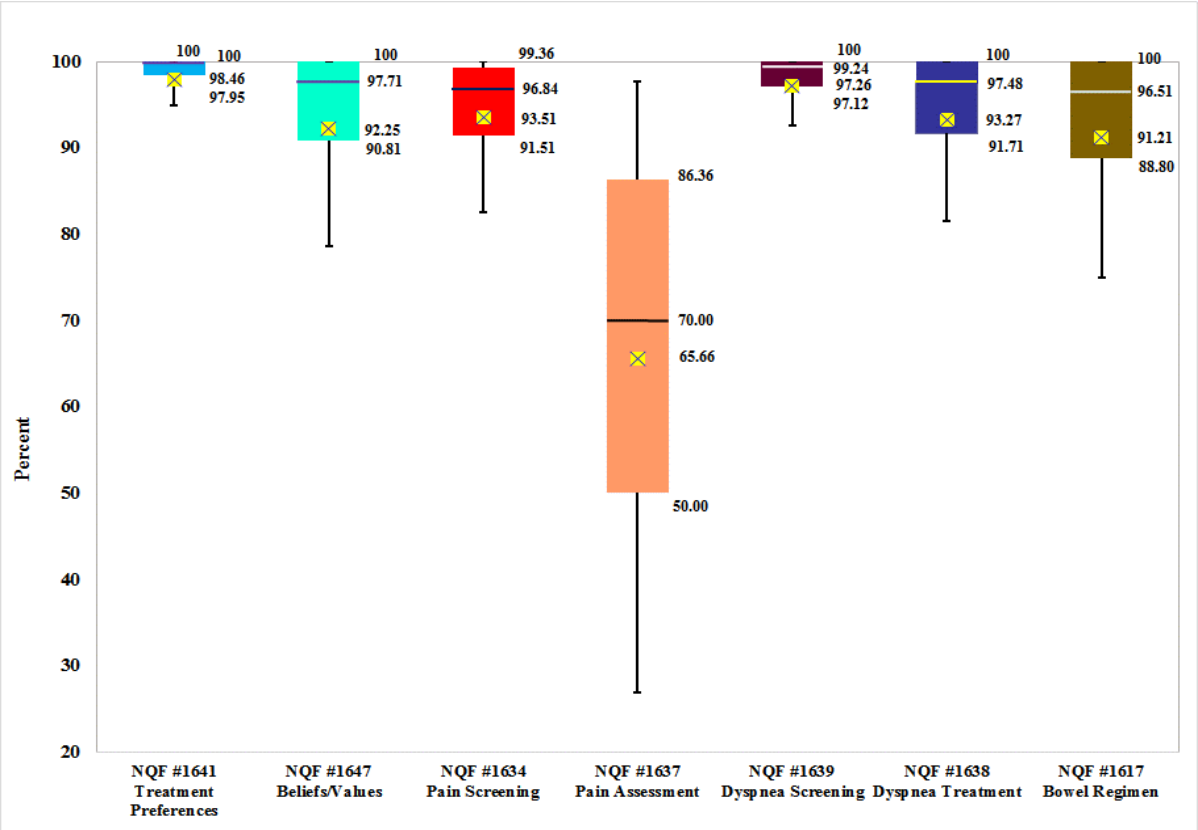
NOTE: N0: Number of hospices with at least one qualifying stay. N: Number of hospices with fewer than 20 qualifying stays.

[2] Removal of the LOS exclusion is under review at NQF as part of comprehensive re-evaluation of the endorsed quality measures.

Distribution of QM Scores and Variability

A meaningful and useful QM should have sufficient variability across providers to distinguish between high- and low-quality hospices. We analyzed the distribution of the hospice-level QM scores to assess the measures' variability. ***Figure ES-2*** presents the distribution of the hospice-level scores on the seven QMs.

Figure ES-2
Distribution of Hospice-Level Scores on HQRP QMs



SOURCE: RTI International analysis of HIS data.

NOTE: The boxes in the figure represent the interquartile range, and horizontal bars within the box represent the median scores. The Xs are the mean scores. The end bars out of the boxes represent the 10th and 90th percentile QM scores. For NQF #1641, Treatment Preferences, the mean was 97.95 percent, slightly lower than the median, 98.46 percent.

- The QM scores were generally high for six of the seven QMs: NQF #1641, Treatment Preferences; NQF #1647, Beliefs/Values; NQF #1634, Pain Screening; NQF #1639, Dyspnea Screening; NQF #1638, Dyspnea Treatment; and NQF #1617, Bowel Regimen. The national mean QM score (represented by Xs in ***Figure ES-2***) ranged from 91.21 percent for NQF #1617, Bowel Regimen, to 97.95 percent for NQF #1641, Treatment Preferences, and the median ranged from 96.51 percent to 100 percent. Compared with these six QMs, hospices' performance on NQF #1637, Pain Assessment, was lower, with a national mean of 65.66 percent and median of 70.00 percent.

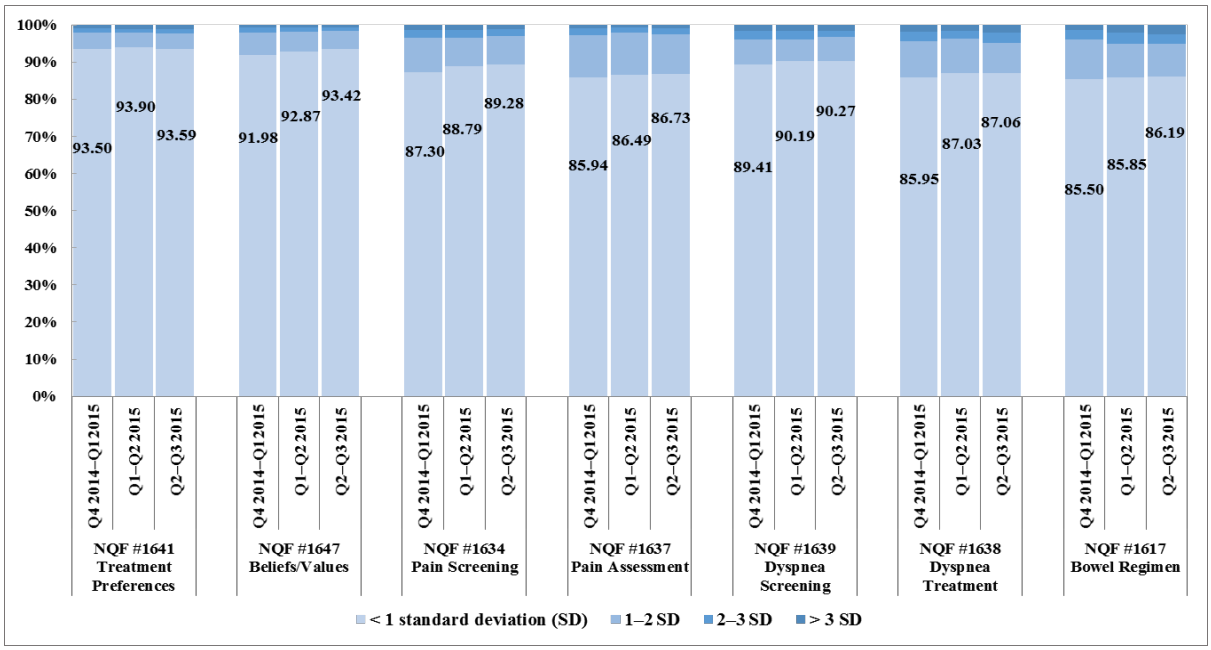
- Six of the seven QMs had relatively small variability, as indicated by the interquartile ranges (the boxes in ***Figure ES-2***). The QM with the smallest variability was NQF #1641, Treatment Preferences, with an interquartile range of 1.54 percent. NQF #1637, Pain Assessment, had the largest variability, with an interquartile range of 36.36 percent.

Reliability

Reliability analyses assess the extent to which a QM produces consistent results about the quality of care. We conducted three sets of analyses to test the reliability of the QMs: stability analysis, split-half analysis, and signal-to-noise ratio analysis.

- *Stability analysis*. Stability analysis describes the extent to which providers' performance assessed by a QM changes across time. ***Figure ES-3*** illustrates the change in facility scores between four consecutive quarters (Q4 2014 and Q1 2015, Q1 and Q2 2015, and Q2 and Q3 2015). The changes in facility scores are reported in standard deviations. Over 85 percent of facilities had a change in QM score of less than one standard deviation, indicating high stability of the QMs. For most QMs, the number of facilities with a change in QM score of less than one standard deviation increased marginally across quarters, suggesting improved reliability over time.

Figure ES-3
Stability of QMs: Standardized Score Change in QM Scores Across Quarters



SOURCE: RTI International analysis of HIS data.

- *Split-half analysis.* Split-half analysis assesses the internal consistency of a QM by randomly dividing the patient stays within each hospice into two halves and calculating the correlation between the hospice's QM scores on the basis of the two randomly divided halves. In this analysis, we conducted a split-half reliability

analysis on all facilities with 20 or more patient stays counted in the measure denominator, and we used the intraclass correlation coefficients (ICCs) to measure the internal reliability. The ICCs of the QMs ranged from 0.82 for NQF #1617, Bowel Regimen, to 0.94 for NQF #1647, Belief/Values, indicating high internal reliability ***(Table ES-2)***.

Table ES-2
Reliability of QMs: Split-Half Reliability and Signal-to-Noise Analysis

QMs	ICC	Signal-to-Noise Ratio
NQF #1641, Treatment Preferences	0.91	0.98
NQF #1647, Beliefs/Values	0.94	0.99
NQF #1634, Pain Screening	0.86	0.97
NQF #1637, Pain Assessment	0.91	0.98
NQF #1639, Dyspnea Screening	0.83	0.98
NQF #1638, Dyspnea Treatment	0.86	0.96
NQF #1617, Bowel Regimen	0.82	0.95

SOURCE: RTI International analysis of HIS data.

- *Signal-to-noise analysis*. If a measure is reliable, then true differences in provider performance should explain a substantial proportion of the variance in QM scores. We conducted an analysis of variance to determine what proportion of total variance in the QM scores was attributable to differences among providers. The results showed that more than 95 percent of the variance in the QMs was attributable to differences among facilities, indicating strong reliability ***(Table ES-2)***.

Validity

Validity refers to whether a QM captures the actual quality of care that it is intended to measure. We conducted two types of analyses to test the validity of the QMs: a correlation analysis of the seven QMs and analyses of the impact of missing data on measure performance.

- *Correlations.* Providers should perform similarly on QMs that reflect similar care processes and performed on similar patient populations. We conducted nonparametric Spearman rank correlation analysis among all hospices' rankings on the seven HQRP QMs. Our results showed that the Spearman correlation coefficients were positive and statistically significant. The significant positive correlations between every pair of QMs indicate that hospices that provided high-quality care in one area also provided high-quality care in other areas at hospice admission. Overall, the correlations between the QMs were low to moderate, ranging from 0.08 to 0.64. The clustering of QM scores, skewed distributions, and low variability across all seven HQRP QMs can affect the level of correlation between QMs.

- *Missing data.* Missing data represent a potential threat to the validity of an HIS item, which in turn may harm the validity of a QM that relies on the item. To assess whether missing data on HIS items might impact the validity of the QMs, we

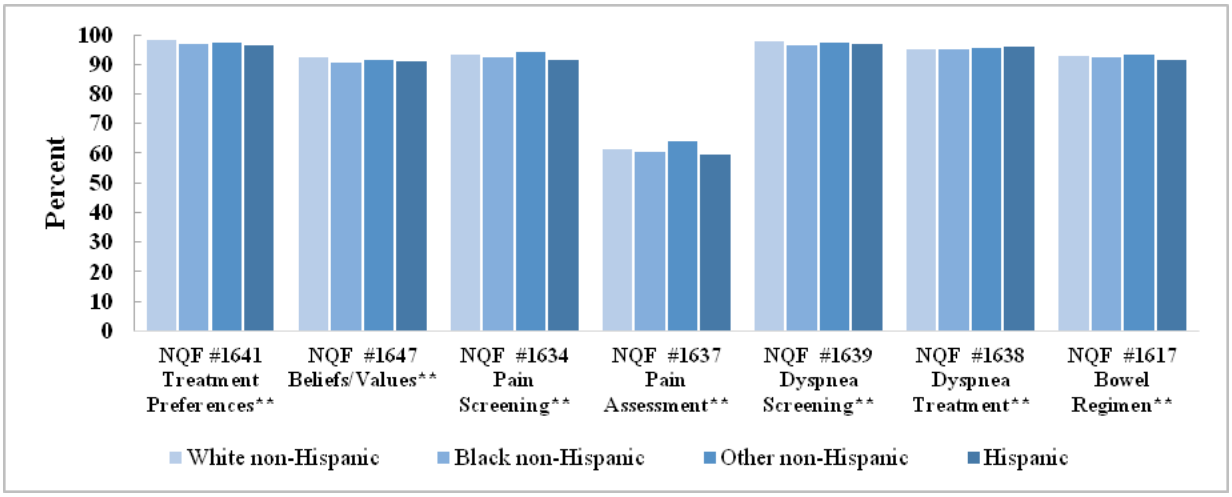
conducted patient stay–level and hospice-level analyses of missingness. For the patient stay–level analysis, we calculated the number and percentage of eligible patient stays for which the HIS-Admission records included a dash for each of the items. The rate of missing data was smaller than 0.2 percent for all items from October 2014 through September 2015. For the hospice-level analysis, we calculated each hospice's percentage of eligible admissions that included missing data for items used to calculate the QM scores. Over 90 percent of hospices did not have any admissions with missing data for these items. These results indicate that missing data for these items have a minimal effect on the validity of the QM.

Disparity Analysis

We conducted a series of disparity analyses at both the patient stay and the hospice level to assess how the measures are affected by the sociodemographic characteristics of hospice patients. At the patient stay level, we compared the QM scores across sociodemographic groups. At the hospice level, we compared the distribution of QM scores between facilities with varying population sociodemographic characteristics.

- *Racial and ethnic disparity.* ***Figure ES-4*** presents the QM scores for the four racial/ethnic groups: white non-Hispanic, black non-Hispanic, other non-Hispanic, and Hispanic. Overall, we found small differences in the QM scores across the four groups. Compared with white non-Hispanic and other non-Hispanic patients, a slightly smaller proportion of the Hispanic and black non-Hispanic patients received the care processes measured by six of the seven QMs. The largest differences were found in two pain-related measures. Hispanic patients received the lowest rates of pain screening and pain assessment (91.30%; 59.74%), followed by black non-Hispanic (92.31%; 60.24%), white non-Hispanic (93.48%; 61.52%), and other non-Hispanic patients (94.28%; 63.87%).

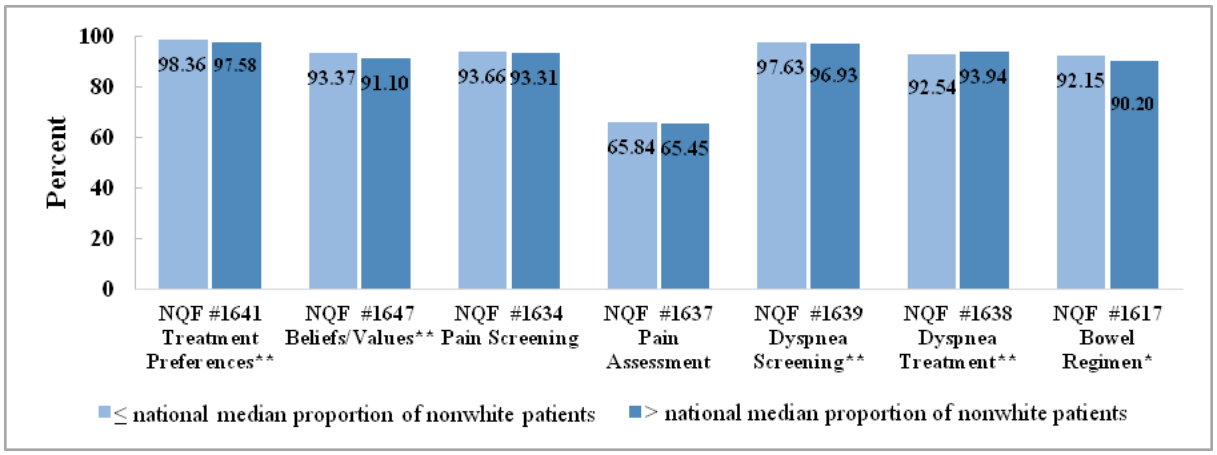
Figure ES-4
Differences in QM Scores Among Four Racial/Ethnic Groups



SOURCE: RTI International analysis of HIS data.
** P < 0.01.

Analyses at the hospice level examined the differences in these measures across two groups: hospices with proportions of nonwhite patients that are less than or equal to the national median, and hospices with proportions of nonwhite patients that are greater than the national median. The findings, presented in ***Figure ES-5***, indicated that for the first group of hospices, the mean hospice-level scores are higher on four QMs: NQF #1641, Treatment Preferences (98.36% vs. 97.58%); NQF #1647, Beliefs/Values (93.37% vs. 91.10%); NQF #1639, Dyspnea Screening (97.63% vs. 96.93%); and NQF #1617, Bowel Regimen (92.15% vs. 90.20%). These hospices had lower scores on NQF #1638, Dyspnea Treatment (92.54% vs. 93.94%). These differences do not seem to be clinically substantial (< 1% for all QMs). Differences in the scores for NQF #1634, Pain Screening, and NQF #1637, Pain Assessment, were not statistically significant.

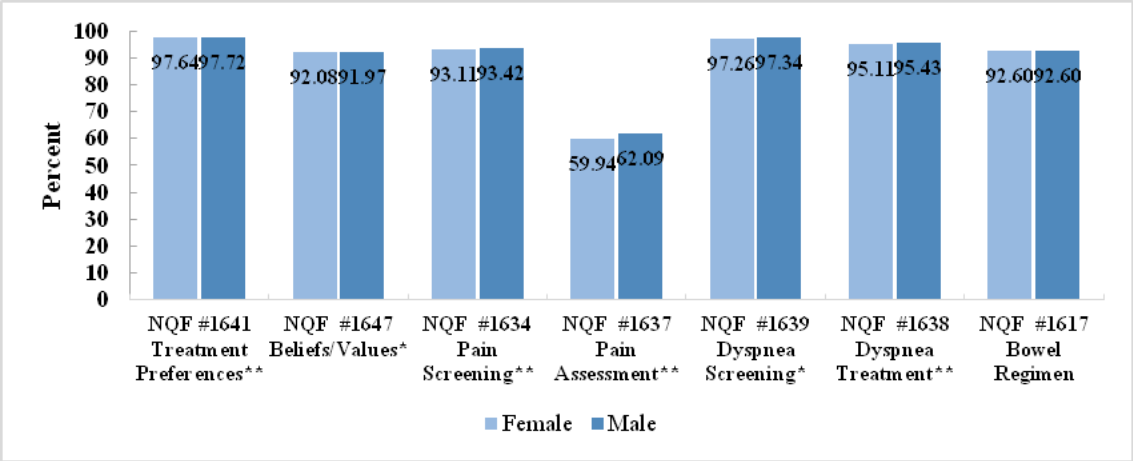
Figure ES-5
Mean Hospice-Level QM Scores by Proportion of Nonwhite Patients



SOURCE: RTI International analysis of HIS data.
* P < 0.05. ** P < 0.01.

- *Gender disparity.* The patient stay–level analysis, presented in ***Figure ES-6***, indicated small gender disparity in some care processes. A smaller portion of female patients received these five care processes: discussion of treatment preferences (97.64% vs. 97.72%), pain screening (93.11% vs. 93.42%), pain assessment (59.94% vs. 62.09%), dyspnea screening (97.26% vs. 97.34%), and dyspnea treatment (95.11% vs. 95.43%). Female patients were asked about beliefs and values more frequently than were male patients (92.08% vs. 91.97%). The average hospice-level scores were not significantly different, statistically or clinically, between hospices with higher proportions of female patients than the national median and hospices with lower proportions.

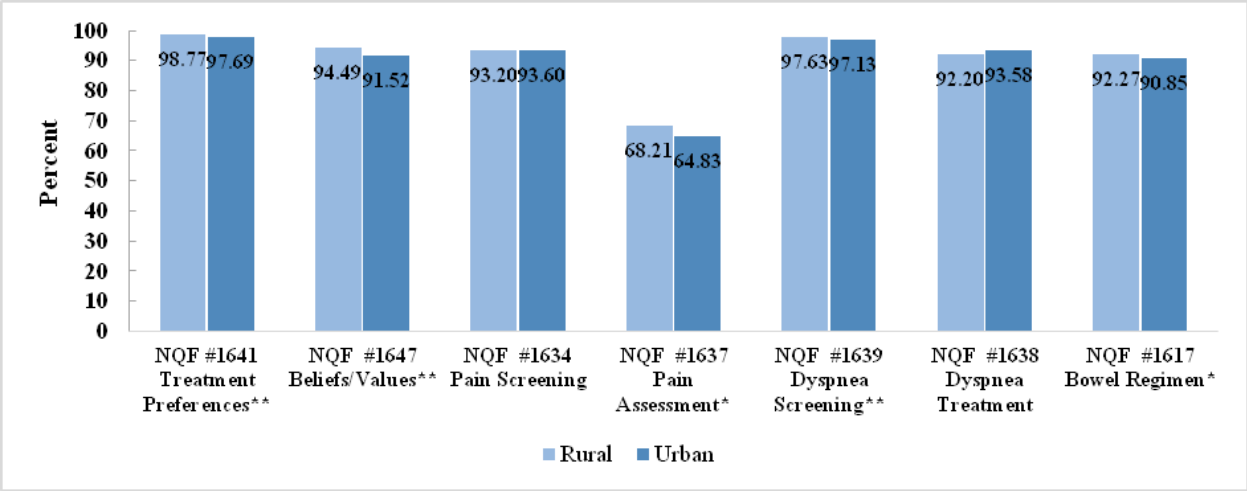
Figure ES-6
Differences in QM Scores Among Female and Male Patients



SOURCE: RTI International analysis of HIS data.
* P < 0.05. ** P < 0.01.

- *Rural-urban disparity.* ***Figure ES-7*** presents the average QM scores of rural and urban hospices. Compared with urban hospices, rural hospices had higher scores on five QMs: NQF #1641, Treatment Preferences (98.77% vs. 97.69%); NQF #1647, Beliefs/Values (94.49% vs. 91.52%); NQF #1637, Pain Assessment (68.21% vs. 64.83%); NQF #1639, Dyspnea Screening (97.63% vs. 97.13%); and NQF #1617, Bowel Regimen (92.27% vs. 90.85%). Rural hospices had lower scores on NQF #1634, Pain Screening (93.20% vs. 93.60), and NQF #1638, Dyspnea Treatment (92.20% vs. 93.58%).

Figure ES-7
Mean Hospice-Level QM Scores in Rural and Urban Areas



SOURCE: RTI International analysis of HIS data.
* P < 0.05. ** P < 0.01.

Summary of Findings

Our analyses demonstrate that the current seven HQRP QMs perform well on various tests of measure performance.

- *Data selection period*. To generate statistically reliable QM scores, we recommend using four quarters of data for measure calculation.

- *LOS exclusion*. Applying the 7-day LOS exclusion significantly reduces the denominator size and excludes some hospices from measure calculation and reporting; the impact of the exclusion on measure performance is minimal, and clinical practice reveals that the care processes captured by these QMs are performed on the day of admission for the vast majority of stays. Therefore, we recommend modifying the current measure specifications by removing the 7-day LOS exclusion to increase denominator size and capture a larger number of hospice stays.

- *Public reporting*. When applying a minimum denominator size of 20 stays, using rolling 12 months of data to create QMs would suppress about 10 percent of hospices from public reporting on four out of the seven QMs, and about 20–30 percent from public reporting on the other three QMs.

- *Distribution and variability*. The overall distribution and variability of the seven QMs indicate that hospices are providing the required and recommended care to the majority of the patients around hospice admission, demonstrating overall high quality of care. There is larger variation across hospices in providing standardized pain assessment to patients who screened positive for pain symptoms. Among the currently assessed and reported care processes in HQRP, pain assessment is the one that can be used to better differentiate hospices based on the quality of care. At the same time, hospices have a larger opportunity to improve their performance in this area.

- *Reliability*. The current seven HQRP QMs have high reliability, demonstrated by high stability, internal consistency, and signal-to-noise ratio.

- *Validity*. Hospices that are high-performing on one care process also provide high-quality care on other care processes around hospice admission. Overall, the correlations between the QMs are low to moderate. Although the correlations of QMs for other care settings are also generally low, the clustering of QM scores, skewed distributions, and low variability across all seven HQRP QMs can affect the level of correlations between QMs. Missing data for HIS items that are used to calculate QM scores have minimal impact on the validity of the QM.

- *Disparities*. We found small differences in QM scores across sociodemographic groups at both patient stay and hospice levels. However, the differences do not seem to be clinically substantial and meaningful.