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: Ctrl + click link 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 #: 1623

Corresponding Measures:

De.2. Measure Title: Bereaved Family Survey

Co.1.1. Measure Steward: Department of Veterans Affairs / Hospice and Palliative Care

De.3. Brief Description of Measure: This measure calculates the proportion of Veteran decedent´s family members who rate overall satisfaction with the Veteran decedent´s end-of-life care in an inpatient setting as "Excellent" versus "Very good", "good", "fair", or "poor".

1b.1. Developer Rationale: The measure is used to improve the quality of end-of-life care received by Veterans and family members during the last month of life in a VA inpatient setting.

S.4. Numerator Statement: The numerator is comprised of completed surveys (at least 12 of 17 structured items completed), where the global item question has an optimal response. The global item question asks "Overall, how would you rate the care that [Veteran] received in the last month of life" and the possible answer choices are: Excellent, Very good, Good, Fair, or Poor. The optimal response is Excellent.

S.6. Denominator Statement: The denominator consists of all inpatient deaths for which a survey was completed (at least 12 of 17 structured items completed), excluding:

1) deaths within 24 hours of admission (unless the Veteran had a previous hospitalization in the last month of life);
2) deaths that occur in the Emergency Department (unless the Veteran had a prior hospitalization of at least 24 hours in the last 31 days of life);

Additional exclusion criteria include:

1) Veterans for whom a family member knowledgeable about their care cannot be identified (determined by the family member´s report); or contacted (no current contacts listed or no valid addresses on file);
2) absence of a working telephone available and valid mailing address to the family member.

S.8. Denominator Exclusions: - Veterans for whom a family member knowledgeable about their care cannot be identified (determined by family member´s report)

- Absence of a current address and/or working telephone number for a family member or emergency contact.
- Deaths within 24 hours of admission without a prior hospitalization of last least 24 hours in the last 31 days of life.
- Deaths that occur in the operating room during an outpatient procedure.
- Deaths due to a suicide or accident
- Surveys in which less than 12 items were answered.

De.1. Measure Type: Outcome: PRO-PM

S.17. Data Source: Instrument-Based Data

S.20. Level of Analysis: Facility, Other

IF Endorsement Maintenance – Original Endorsement Date: Feb 14, 2012 **Most Recent Endorsement Date:** Jan 07, 2015

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? N/A

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.

Evidence Summary

The developer provides a logic model stating that receiving a palliative care consult or dying in a hospice unit results in a greater likelihood of families rating end of life inpatient care as excellent.

The developer states that physical symptoms such as pain, nausea, constipation, and dyspnea are common at end of life and that clinicians do not always recognize these symptoms or manage them appropriately. The developer states that studies have found that providers do not communicate with patients about patients' health care preferences and that providers' treatment decisions may not be consistent with patients' preferences.

The BFS has consistently shown significant associations (p<0.001) using logistic/linear regression tests with quality of care indicators based on the empirical literature and "Best Practices" as outlined in the National Consensus Project for Quality Palliative Care Clinical Guideline (presence of a palliative consult at death, death in an inpatient hospice unit, chaplain and bereavement contacts with patients and family members).

The developers demonstrate in their application that nonresponse and patient case-mix adjusted facility-level BFS-PM scores are consistently higher for when patients receive these quality indicators. Weighted linear regression analyses demonstrate statistically significant, positive associations between receipt of a quality indicator and facility-level BFS Performance Measure scores.

The developer included a recommendation from the 2009 version of the *Clinical Practice Guidelines for Quality Palliative Care*. The specific recommendation is:

- Guideline 7.1 Signs and symptoms of impending death are recognized and communicated in developmentally appropriate language for children and patients with cognitive disabilities with respect to family preferences. Care appropriate for this phase of illness is provided to patient and family.

In the disparities section of the measure submission, the developer shared results of an analysis demonstrating higher scores for five of six domains on the Family Assessment of Treatment at the End of life (FATE) survey for patients who received a palliative consultation.

Changes to evidence from last review

☒ **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:**

Updates:

Question for the Committee:

- ○ *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- ○ *This measure is derived from family report, does the target population value the measured outcome and find it meaningful?*

Guidance from the Evidence Algorithm

Patient reported outcome (Box 2) à Relationship demonstrated between an outcome and a process à 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.

In the testing attachment, the developers provide figures demonstrating significant variation in BFS-PM scores across VA facilities nationwide and over time. The developer also includes citations to a number of published articles that have documented statistically and clinically significant differences in performance with regard to:

1) receipt of a palliative consult;
2) death in an inpatient hospice unit vs. other inpatient venues;
3) diagnoses at the end of life;
4) gender;
5) receipt of aggressive care;
6) chaplain interaction;
7) bereavement support after death;
8) DNR order at time of death;
9) race/ethnicity;
10) family involvement; and
11) timing of palliative consults.

In the testing attachment, the developer provides results from 2017 (n=146 facilities) demonstrating a 65% mean overall score, a score range from 13% - 100%, and IQR of 85 and 72.

Disparities
The developer has documented the presence of significant racial/ethnic disparities on the BFS-Performance Measure in previously published work. The developer also shares analysis of Family Assessment of Treatment at the End of life (FATE) results that indicate ethnicity (white vs. nonwhite) and older age were independently associated with higher FATE scores.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?

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 to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.

- N/A
- Bereaved Family Survey (1623) is a composite survey measuring the proportion of Veteran decedent´s family members who rate overall satisfaction with the Veteran decedent´s end-of-life care in an inpatient setting, an outcome measure indicative of satisfaction with the care received. I am not aware of additional studies. I am concerned that there is not opportunity for the seriously ill person to indicate their satisfaction with care received (a general concern with surveying family of a decedent weeks after the stay).
- N/A
- NA
- Instructed not to answer
- skip
- n/a
- NA
- Yes
- This measure is still not strong in demonstrating outcomes but I can support continuation
- Although family report is less ideal than patient self-report, that is not possible for patients after death. Therefore, this seems the next best thing. While not perfect, this survey gets at key aspects of end-of-life care and has been used to improve care within the VA system.
- The developers make the case that a palliative care consult and/or inpatient hospice results in an increased satisfaction of families for EOL care resulting in higher family scores. This indicates that the family values the measure and finds it meaningful.

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- N/A
- The developer provides performance data from 2017 (n=146 facilities). The developer has documented the presence of significant racial/ethnic disparities on the BFS-Performance Measure in previously published work. Disparities continue to exist with the developer also sharing an analysis of Family Assessment of Treatment at the End of life (FATE) results that indicate ethnicity (white vs. nonwhite) and older age were independently associated with higher FATE scores. and over time,
- N/A
- NA
- Instructed not to answer
- skip
- n/a
- NA
- Performance gap noted. Could be used to improve quality.
- Criteria met to continue.
- Yes, there is there a gap in care that warrants a national performance measure
- In previously published work the developer documented significant racial/ethnic disparities. They have demonstrated statistically and clinically significances in performance based on 11 different criteria which indicates room for improvement.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

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.

Methods Panel Review (Combined)

[Fall 2020 Evaluation Cycle] Methods Panel Evaluation Summary:

- **Ratings for reliability:** H-1; M-6; L-1; I-0 → Measure passes with MODERATE rating
- Reliability testing conducted at the measure score and data element level:
  - To demonstrate reliability of the survey item used in this measure, the developers conducted four test-retest analyses on 93 randomly selected Bereaved Family Survey (BFS) respondents who agreed to complete the BFS on a second occasion (30 days apart)
    - Analysis #1 (Cohen's kappa): Kappa=0.5 (n=92); Developer cites Cohen's article that says a kappa of 0.5 indicates moderate agreement
    - Analysis #2: two-way random effects, absolute agreement, single rater/measurement: ICC (2,1) =0.52 (moderate agreement, according to Cohen)
    - Analysis #3, Logistic Regression: Compared to those who reported BFS=0 at time 1, respondents who reported BFS=1 at time 1 had 17.2 the odds of reporting BFS=1 at time 2 (interpreted as very strong association)
    - Analysis #4, Cohen's d Effect Size of a 2x2 contingency table: d=1.57 ("large" effect when d≥0.8)
  - Developers also described an analysis of the global item obtained via phone vs. mail administration (2009-2012 data for phone, 2012-2017 data for mail). Results indicate both are normally distributed (mean=58, 63 respectively, both with a standard deviation of 5), and very few facilities had mean ≥ 90, interpreted as no ceiling effects. Developers also reported Cronbach's alpha for phone vs. mail (0.81 vs 0.83).
  - To demonstrate reliability of the measure score, the developers conducted two analyses:

        - ICC1 using a mixed-effects logistic regression model:
            - FY10-FY12 (administered predominantly as a phone survey)
                - Facility-level variance estimate=0.15; 95% CI .12-.20; p<0.001
                - ICC1=0.04 (95% CI: .03-.06)
            - FY13-FY17 (administered predominantly via a mail survey):
                - Facility-level variance estimate =0.13; 95% CI .09-.20; p<0.001.
                - ICC1=0.04 (95% CI: .03-.06)
        - Split-half analysis with application of Spearman-Brown prophecy formula: 0.89
        - SMP members commented that reliability at the data element level is marginal and reliability at the measure score level is acceptable, but the reported ICC value of .04 is low. Because of the stronger measure score level testing findings, SMP members passed the measure on reliability.
- **Ratings for validity:** H-0; M-7; L-0; I-1 → Measure passes with MODERATE rating
- Validity testing conducted at the measure score and data element level:
    - To demonstrate validity of the survey item used in this measure, the developers analyzed five percent (randomly selected) of written responses to the question, “Is there anything else that you would like to share about the Veteran’s care during the last month of life?” These comments were categorized as positive, neutral, or negative. These categorizations were correlated with the responses from the overall rating of care item (the item from the survey used in this measure).
        - Spearman correlation coefficient=0.51; p<0.001
    - Using patient-level data (N=84,616) and facility-level data (N=146), the developer ran nine separate logistic/linear regressions adjusted for nonresponse bias and patient case-mix. The independent variables were the process measures and the outcome variable was the individual BFS item and facility and patient level BFS percent “excellent.” Their hypothesis was that receipt of each of the “best practices” processes should result in a statistically significant higher BFS score and the results of these analyses support the developer’s hypotheses. Logistic regression analyses demonstrate statistically significant, positive associations between receipt of quality indicators, and patient-level BFS Performance Measure scores.
    - In the analysis of exclusions/missing data, a total of 16 percent of eligible decedent veterans were excluded from the measure. A total of 4 percent were excluded because they died within 24 hours of admission. The remaining excluded cases were included in a nonresponse bias analysis.
    - Prior to reporting of facility-level scores, the BFS-Performance Measure is adjusted for patient case-mix and survey nonresponse, and are stratified by facility complexity level. The measure is risk-adjusted using five factors (veteran’s age at the time of death; number of medical comorbidities present at the time of death; veteran’s primary diagnosis on last admission; relationship of veteran’s next-of-kin (i.e., spouse), and model of administration mode (i.e., mail).

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 had concerns about the lack of social risk factors included in the risk-adjustment. Do you share these concerns?
- The Scientific Methods Panel was not 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**

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

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- N/A
- Reliability at the data element level is marginal and reliability at the measure score level is acceptable, but the reported ICC value of .04 is low. Because of the stronger measure score level testing findings, SMP members passed the measure on reliability.
- N/A
- NA
- Instructed not to answer
- skip
- n/a
- NA
- No concerns
- It is limited and yet can support
- I have no concerns on this point.
- It appears that this measure can be consistently implemented.

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

- N/A
- data element reliability is marginal but at the composite measure acceptable
- N/A
- NA
- Instructed not to answer
- skip
- n/a
- NA
- No

- limited concern
- I have no concerns on this point.
- no concerns discussion is not necessary

2b1. Validity -Testing: Do you have any concerns with the testing results?

- N/A
- No concerns with testing results noted.
- N/A
- NA
- Instructed not to answer
- skip
- n/a
- NA
- No
- limited
- I have no concerns on this point.
- I am concerned about the lack of social risk factors and the use of randomly selected written response to demonstrate validity is concerning.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- N/A
- None noted
- N/A
- NA
- Instructed not to answer
- skip
- n/a
- NA
- Question if only family are contacted vs. primary contact or caregiver
- adequate
- I have no concerns on this point.
- Exclusions seem to be consistent, there is a lack of social risk factors in the risk adjustment that is concerning.

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- N/A

- Missing data does not constitute a threat to the existing measure but the construct excludes the person who received the care when possible, a significant issue related to validity of construct.
- N/A
- NA
- Instructed not to answer
- skip
- n/a
- NA
- No
- outcome data would be helpful but difficult to obtain
- I have no concerns on this point.
- There does not appear to be any missing data, statistically significant differences, or multiple data sources.

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.

- Some variables (death, risk-adjustment) are available electronically from defined fields in the EHR.
- Other variables are obtained via mailed/telephone survey. There is no fee or licensing to use the measure as specified.
- The developer states they have been able to refine both procedures for gathering electronic data and survey contact procedures for more efficient survey administration.

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: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?

- N/A
- No concerns noted.
- N/A
- NA
- Instructed not to answer

- skip
- n/a
- NA
- No concerns
- limited consistent collection
- I have no concerns on this point.
- Some of the required data is available via EHR such as death and risk adjustment the others are obtained via a mailed or telephone survey.

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 developer's discusses plan for starting to administer the survey. The developer states once they administer the survey, they plan to distribute results within the VA to:

1) Assist VA in selecting community nursing homes for veterans' end of life care
2) Provide community comparisons on the quality of end of life care in VA nursing homes and
3) Give veterans and their families meaningful quality information regarding end of life care

The developer provided the following additional details since the 2019 submission:

- BFS is not currently used for consumer choice, but for transparency and quality improvement efforts.
- BFS results are reported on an internal VA dashboard which promotes accountability
- Developers plan to modify the scaling to allow for comparisons between BFS-PM and CAHPS-Hospice for community nursing home settings prior to the next submission.

- BFS has been adopted by non-VA organizations. Stanford, Duke, UCLA, and Kaiser medical centers are all using the BFS to guide quality improvement efforts.

At the previous endorsement in 2015, the measure developer expressed confidence that the measure would be publicly reported by the next endorsement cycle (now).

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 states that survey results are reported to VA stakeholders on a quarterly (or as-needed) basis. Stakeholders receive the performance measure results for this measure and the other survey results as well. The VA Hospice and Palliative Care Implementation Center and the VA Veteran Experience Center have regular contact with stakeholders to assist with interpreting results and data. A SharePoint for best practices has been created to assist with quality initiatives. Feedback on the measure is obtained through regular conference calls with stakeholders.

Additional Feedback: None included.

Questions for the Committee:

- How has the measure been vetted in real-world settings by those being measured or others?

Preliminary rating for Use: ☐ Pass ☒ No Pass

RATIONALE: No examples of public reporting or use in accountability programs are provided. The measure has been endorsed since 2012, which places it past the three-year accountability requirement and the six-year public reporting requirement for maintenance measures.

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 states, “National patient-level improvement is demonstrated by increasing scores over time from 52% N=8,432 in 2008 to 64% N=10,899 in 2018. Facility scores vary over time with 79% N=116 out of 142 facilities having an increase in scores between 2008 and 2018. In 2018, 81 facilities (55%) scored above the national mean of 64%.”

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 reports family member issues were identified through contact for survey completion (suicidal thoughts, need for bereavement services, lack of information about benefits, etc.) and that they were able to assist with addressing these issues. These were unexpected benefits of implementing the survey.

Potential harms

No potential harms described.

Additional Feedback:

None.

Questions for the Committee:

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Are you aware of any potential unintended consequences not mentioned here?

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

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

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- This measure is publicly reported in the VA even outside of the organization that is being measured. The VA also publishes widely the results of their findings (e.g. https://pubmed.ncbi.nlm.nih.gov/25793359/). I have zero concerns regarding its use.
- The VA has adopted the measure and uses to guide performance improvement. Additionally, BFS has been adopted by non-VA organizations. Stanford, Duke, UCLA, and Kaiser medical centers are all using the BFS to guide quality improvement efforts.
- N/A
- Not publicly reported. Used within the VA as part of their accountability program. The measure developer notes feedback has been obtained
- In reading the developer's report, it appears that the measure has been vetted in real-world settings by those being measured or others. The results are reported to the VA stakeholders and contact is made with the stakeholders to ensure they understand the results. Also, the VA obtains feedback on the survey via regular conference calls. I am not sure that the feedback from stakeholders has been considered with regard to changes, if there have been any.
- Survey results are reported to VA stakeholders on a quarterly basis and they receive performance measure results. A Share Point for best practices has been created.
- Performance results are disclosed and available outside of the specific practices whose performance is measured and efforts are underway to enhance transparency and report directly to veterans and larger public. Feedback on measure is requested.
- Facility/VISN performance on BFS is “publicly” reported within VA as each facility/VISN/VA leadership has access to all results (available via Hospice and Palliative Care Data Dashboard link as well as housed in

the VA's internal databases) in addition to the public reporting that occurs regularly in academic journals. Developer has not been able to obtain needed governmental approvals for widespread public reporting. Survey results reported to VA stakeholders quarterly. Feedback on measures is obtained through regular conference calls with stakeholders.

- Unclear how measure is publicly viewed. Report says it is publicly reported in VA system.
- limited
- I have no concerns on this point.
- The measure is reported in the VA network but not publicly. There is an accountability program in use. There is feedback on the measure via several mechanisms including conference calls and quality initiatives.

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- N/A
- Identification and the ability to address suicidal ideation noted. No harms noted.
- No concerns
- Results will be utilized for quality improvement. No harms noted
- Instructed not to answer
- The benefits of improving care at the end of life outweighs any unintended negative consequence.
- Moderate; Usability for improvement is highlighted by description of how results are shared within VA systems, how non-VA systems have adapted the measure for improvement of care in their organizations, and unintended benefits realized. Further information regarding potential harms would be helpful.
- NA
- None
- The measure allows for focus on this important service
- I have no concerns on this point.
- No potential harms were indicated. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare outweigh evidence of unintended negative consequences.

Criterion 5: Related and Competing Measures

Related or competing measures

- 2651: CAHPS® Hospice Survey (experience with care)

Harmonization

The developer states that the populations are different for these two measures, as 1623 is focused on deaths in a VA inpatient setting.

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

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- 2651: CAHPS® Hospice Survey (experience with care) is a related and competing measure. The VA will begin comparison of data from both measures and provide in next endorsement cycle for BFS.
- N/A
- CAHPS HOSPICE SURVEY however a different population of patients as this measure focused on deaths within the VA inpatient setting
- Instructed not to answer
- NA
- This is used only for veterans.
- none of which I am aware
- I have no concerns on this point.
- No competing measures

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/26/2021

- No NQF Members have submitted support/non-support choices as of this date.
- No Public or NQF Member comments submitted as of this date.

Combined Methods Panel Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 1623

Measure Title: Bereaved Family Survey

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.*

2. **Briefly summarize any concerns about the measure specifications.**

   Panel Member #1: No concerns.

   Panel Member #2: None.

   Panel Member #3: none

   Panel Member #4: No concerns.

   Panel Member #5: No concerns.

   Panel Member #6: I have no concerns. The measure specifications are clear and concise.

   Panel Member #7: None

   Panel Member #8: This PRO-PM is mainly based on the answer to one global survey item (Q18), however, it requires that respondents need to answer at least 12 of 17 survey items (Q1-Q17) to be included. The rationale for this exclusion criterion is to reduce the amount of miss data, however, answers to Q1-Q17 are not used for this PRO-PM. It is not clear if this criterion is needed at all. More importantly, missing values on Q1 – Q17 may potentially be related to answer to Q18, so this unnecessary exclusion may lead to potential bias.

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 VALIDITY testing** of **patient-level data** conducted?

   ☐ **Yes** ☐ **No**

   Panel Member #1: N/A

   Panel Member #2: None.

   Panel Member #3: none

   Panel Member #5: NA – score & data element testing conducted

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   Panel Member #1: Methods used for reliability testing are appropriate and reasonable.

   Panel Member #2: None.

   Panel Member #3: Item-level reliability focused on assessing agreement between responses provided by the same respondent on two different occasions for the question on which the measure is based. These analyses were based on data from 93 respondents who agreed to repeat the survey. Four different metrics were used to quantify reliability.

   Developers used two methods (random effects modeling and split half testing) to estimate patient-level intraclass correlation coefficients (ICCs) and used the Spearman-Brown prophecy formula to determine the number of respondents per provider that would be required in order for provider-specific average scores to have at least 0.70 reliability.

   Panel Member #4: Data element testing included: BFS test retest reliability; test retest reliability (Cohen's Kappa); test retest reliability – two way random effect single rater agreement; test retest reliability logistic regression; test retest Cohen's d; and internal consistency. Facility level reliability testing included: interclass correlation; and Spearman Brown split half reliability. These seem appropriate to me.

   Panel Member #5: Note the "Data element reliability testing" heading appears to have some tests below the heading that are score level testing. The testing appears to be appropriate for the given measure.

   Data Element Reliability Testing

- BFS Test-Retest Reliability
- Test-Retest Reliability of Absolute Agreement, Statistical Analysis #1 (Cohen's kappa)

- Test-Retest Reliability of Absolute Agreement, Statistical Analysis #2: two-way RANDOM effects, absolute agreement, single rater/measurement
- Test-Retest Reliability, Statistical Analysis #3, Logistic Regression
- Test-Retest Reliability, Statistical Analysis #4, Cohen's *d* Effect Size of a 2x2 contingency table' [p6]

Data element reliability testing (Internal Consistency Reliability):
Facility-level Reliability Testing I (Intraclass Correlation Coefficient):
Facility-level Reliability Testing II (Spearman-Brown Split-Half Reliability)' [p7]

Panel Member #6: Test-retest reliability was assessed, with consecutive tests conducted 30 days apart, using an array of statistical methods, including Cohen's kappa, random effects regression (mixed modeling), logistic regression, and Cohen's *d*. Furthermore, facility-level reliability was assessed, using intraclass correlation coefficient estimation and Pearson-Bowman split-half reliability estimation.

Panel Member #7: Test-retest for patient level and ICCs and Spearman-Brown coefficients (for split-half reliability) for facility-level reliability are appropriate.

Panel Member #8: The developer used a random sample of 93 respondents to assess the test-retest reliability of the global survey item in question. This was to test the repeatability of this item if respondents would provide the same answer on separate occasions. The developer calculated Kappa, ICC, predictive ability of prior answer and Cohen's d.

The description of facility level score reliability testing is somewhat confusing. The minimum sample of 56 should be based on estimated between facilities variance, which the developer should have described first.

Panel Member #9: Many of the reliability test methods used were appropriate, but some were either not appropriate (comparing modes of administration and calling it "internal consistency reliability" or not very compelling or stringent (the logistic regression or Cohen's d effect size estimates for data element reliability).

7. **Assess the results of reliability testing**

Submission document: Testing attachment, section 2a2.3

Panel Member #1: Acceptable results for data element and facility level (score) testing.

Panel Member #3: The item-level reliability results indicate imperfect test-retest reliability. This does not strike me as a concern, however. Low test-retest reliability implies that a larger number of respondents per participant are needed in order to precisely estimate between–provider differences. The score-level reliability results suggest that a sample size of 56 respondents per participant will yield score-level reliabilities above 0.7The developers noted that over 97% of facilities in their testing sample had sample sizes above the required threshold. The average facility-specific reliability estimate was 0.89.

Panel Member #4: Test-Retest Reliability Measures: All measures of test-retest reliability are reported in Table XXX below

1. Cohen's kappa for the 92 respondents measured on two occasions is 0.5. According to Cohen's original article, a kappa of 0.5 indicates "moderate" agreement between time 1 and time 2 BFS assessments.
2. ICC(2,1): The 2-way random-effects, absolute agreement, model estimated an ICC(2,1) = 0.52. This is considered a "moderate" level of agreement according to Cohen's original article.
3. Odds Ratio estimate: Compared to those who reported BFS=0 at time 1, respondents who reported BFS=1 at time 1 had 17.2 the odds of reporting BFS=1 at time 2. An odds ratio of this magnitude represents a very strong association between BFS overall scores at time 1 and time 2.
4. A Cohen's d of 0.8 or greater is considered a 'large' effect size. Our reported Cohen's d of 1.57, therefore, represents a very strong effect of BFS overall score at time 1 on BFS overall score at time 2.

Table XXX. Measures of test-retest reliability between BFS item scores from time 1 and time 2.

Measure	Measures of Test-Retest Reliability	"Rule of thumb" Interpretation
1. Cohens kappa[1]	0.498 (0.26 - .71)	Moderate agreement
2. ICC(2,1)[2]	0.52 (0.36 – 0.66)	Moderate agreement
3. Odds Ratio Estimate[3]	OR = 17.3 (95% CI: 4.7 – 63.3)	Very strong association
4. Cohen's d[4]	1.57	>= 0.8: 'large' effect size

While these numbers are acceptable some are only moderate and the lower bounds of the ranges are rather low.

Panel Member #5:

Test-Retest Reliability Measures:

- Cohen's kappa: 0.5. [moderate]
- ICC: … 0.52. [moderate]
- Odds Ratio estimate: … respondents who reported BFS=1 at time 1 had 17.2 the odds of reporting BFS=1 at time 2. [strong]
- Cohen's d of 1.57 [strong]' [p8]

Panel Member #6: In 92 respondents, Cohen's kappa was 0.50; random effects regression elicited an intraclass correlation coefficient of 0.52, the odds ratio of an excellent rating in the second test for an excellent vs. non-excellent response in the first test was 17; and Cohen's *d* was 1.6.

The facility-level intraclass correlation coefficient was 0.04.

Panel Member #7: Kappa and ICCs for patient level test-retest reliability are moderate. However, ICCs for between vs. within facility variation are low (ICCs= 0.04).

Panel Member #8: The results of data element reliability testing show moderate reliability, kappa and ICC are consistent. OR derived from logistic regression and Cohen's d also provide additional support.

Two steps described in the facility level score reliability testing are actually connected. The developer separated them into two different steps which might cause confusions. ICC reported on 2a2.3 should be derived from a mixed effect model, based on the estimated ICC, the developer would then estimate reliability for each facility after accounting for the sample size of each facility. The reported results are acceptable although the developer could have described their work more clearly.

Panel Member #9: Reliability testing at the data element level yielded marginal results – right at the cusp of "moderate" and "low". Rates of absolute agreement in the relatively small test-retest reliability study were OK – rates reported for kappa and ICC were marginal at best. The developers report a Cronbach's alpha statistic of .-81 and .83 for the mail version of the survey at two time points, but it's not clear how this statistic is calculated or should be interpreted on a single-item scale. Reliability statistics for measure score reliability were better using the Spearman-Brown approach, but it seems a bit concerning that only 4% of the facility-level variance seems due to some hospital-level "signal" from this measure.

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**

   ☐ **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**

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

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

    ☒☐ **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 specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate **INSUFFICIENT** if you believe you do not have the information you need to make a rating decision)

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: Cohen's Kappa and ICC is at the lower end of moderate.

    Panel Member #3: I assigned a "high" rating because the developers report an average estimated signal to noise reliability of 0.89 for the facility-level mean scores. Also, the developers estimate that a sample size of 56 respondents per facility would yield an estimated reliability of 0.70 and they note that 97% of facilities have at least this sample size. Item-level reliability was less impressive but the high facility-level reliability estimates make the item-level results somewhat irrelevant.

    Panel Member #4: No concerns.

Panel Member #5: The salient reliability score level test results are between moderate and strong. My rating of 'moderate' based on the types of tests that performed 'moderate' and 'strong'.

Panel Member #6: The data convincingly indicate that test-retest reliability is very high, but the facility-level reliability appears to be low.

Panel Member #7: Although patient level reliability estimates are adequate, the within facility variation appears to be high relative to the between facility variation as reflected in ICCs of 0.04.

Panel Member #8: The developer conducted appropriate reliability testing for both data element and performance score. Results are acceptable for both. The developer could have described their approach on facility level score testing more clearly.

Panel Member #9: As noted earlier, reliability at the data element level is marginal, and reliability at the measure score level is better, but the reported ICC value of .04 is not impressive. Since the distinction between "moderate" and "low" is essentially pass vs. fail, I chose "pass", but only on the basis of the measure score reliability findings. It's a close call.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

Submission document: Testing attachment, section 2b2. No concerns.

Panel Member #3: None

Panel Member #4: No concerns.

Panel Member #5: 3 concerns:

[1] MIF lists the following exclusions, that are not mentioned in the exclusion section in the testing form

- Deaths that occur in the operating room during an outpatient procedure
- Deaths due to a suicide or accident

Because these exclusions are not noted in the testing form, there are no findings regarding the exclusions

[2] Of the exclusions in the testing form, it results in 16% of cases being excluded, which is relatively high.

[3] Of the exclusions in the testing performed does not do a very good job of directly identifying the impact of the exclusions and the degree to which they do / don't impact scoring / survey results.

Panel Member #6: Survey nonresponse is influential.

Panel Member #7: None

Panel Member #8: Excluding surveys without answers to at least 12 of 17 survey items is not necessary and may even potentially lead to bias.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

Submission document: Testing attachment, section 2b4. No concerns2b5.

Panel Member #1: No concerns.

Panel Member #4: No concerns.

Panel Member #5: The findings do not directly present results of the differences in findings from the instrument. The figure presented in response to 2b4 comes the closest. However, there are issues with the figure:

[1] It presents scores of percent of "rated overall end of life care as 'excellent'". We don't know if that is the only, primary or 1 of many results produced / reported via the instrument.

[2] In the title of the figure it states "FY10 – FY17". So it appears we're not getting results from the most recent administration of the survey, but approximately 8 years of survey results presented, which one is unable to discern what year is the most recent.

[3] The findings presented don't clearly answer the question of to what degree does the measure yield meaningful differences in provider performance.

The response to 2b4.2 points to cites in response to 2b4.1. However, the summaries provided of these cites don't answer the question of to what degree does the measure yield meaningful differences in provider performance.

Panel Member #6: There is clear variation in the measure among facilities, but the measure steward presents no data regarding circumstances in which one facility may be statistically distinguished from another. Nonetheless, the caterpillar plot is a compelling visual.

Panel Member #7: Because the within facility variations appears to be substantial, the ability to discriminate between individual facilities does not appear to be supported by the data. However, there appear to be significant differences between the highest and lowest quartiles (see 2b4.1, p. 23).

Panel Member #8: No concerns

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. None.

Panel Member #4: No concerns

Panel Member #5: No concerns

Panel Member #6: I have no concerns.

Panel Member #7: None

Panel Member #8: No concerns

15. **Please describe any concerns you have regarding missing data.**

Panel Member #1: None

Panel Member #2: None.

Panel Member #3: A limitation that applies to all voluntary surveys is the potential for nonresponse bias. The developers state that they adjusted for nonresponse bias but the methods are not described.

Panel Member #4: I do not have concerns about missing data per se – but rather about response rates as they seem to be widely variable from facility to facility (29% to 73%). This seems like it could produce biased data.

Panel Member #5: No concerns

Panel Member #6: There is substantial nonresponse, and the potential for absolute changes of 10 percentage points in the measure exists. I do not understand how this measure is used, so I do not know whether measure changes between 5 and 10 percentage points are potentially impactful.

Panel Member #7: Although response rates are low, they are in the range observed by other survey based measures. Weighting for missing values did not change mean facility scores substantially.

Panel Member #8: No concern. As this is really based on answer to one global survey item.

Panel Member #9: The developers point out problems with non-response bias and modest survey response levels that vary by type of facility, leading to some potential problems with interpretation of between-facility differences. There is an adjustment for non-response in the adjustment model, so they have done what is possible to do with this problem**.**

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? 2b3.3a ☐ Yes ☐☒☒ No ☒☐ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☒☒☒☒☒2b3.3a ☐ Yes ☐☒ No Rational for non-inclusion is present.

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

Panel Member #1: No concerns.

Panel Member #5: NA – Not fully addressed

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒☐☒ Yes ☐ No Strictly speaking, length of hospice stay is a post-treatment variable. I think it's okay to gloss over this.

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ 2b3.3a, 2b3.3b. ☒ Yes ☐ No N/A

Panel Member #5: NA – factors present at start of care

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ 2b3.4b, 2b3.5 ☒ Yes ☒☒☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒☒☒☒☒2b3.6, 2b3.7, 2b3.8, 2b3.9, 2b3.10☐ Yes ☐☒ No

16d.5.Appropriate risk-adjustment strategy included in the measure? ☒☐ Yes ☒☒☐ No

Panel Member #3: I am unclear about this – see comments below.

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

Panel Member #1: The approach taken by the measure developer is consistent with other like surveys and the VA's policy on adjusting for social determinants of health**.**

Panel Member #3: The developers describe the adjustment method as follows: " Facility-level scores were adjusted for case mix using inverse probability weighting. First, all case-mix adjustor variables were entered into a logistic regression model, predicting a response of "excellent" on the BFS-PM at the patient level. A propensity score, or predicted probability, for an "excellent" response was derived from the results of the logistic regression. Finally, weights for the case-mix adjustment were calculated by taking the reciprocal of the propensity score and were applied all facility-level BFS outcomes."

I am struggling to see how weighting responses by the inverse of the predicted probability of an excellent response would provide an appropriate adjustment for case mix across providers. Were the methods perhaps described incorrectly? An alternative approach would involve weighting responses by the inverse of the probability of receiving care from the particular hospice that is being assessed.

Panel Member #4: The developers state "We examined 5 variables as potential case-mix adjustors: Veteran's age at the time of death (in years), number of medical comorbidities present at the time of death as defined by Van Walraven and colleagues' modification of the Elixhauser score, Veteran's primary diagnosis on last admission (classified into 1 of 15 clinical categories using the Agency for Healthcare Research and Quality Clinical Classification Software), relationship of Veteran's next-of-kin (i.e., spouse), and BFS administration mode (i.e., mail)." AND "Social risk variables such as race and ethnicity were not included in the model because the BFS is used as a quality improvement measure. If you adjust for race/ethnicity (knowing that disparities exist), you are essentially letting facilities "off the hook" if they care for more racial/ethnic minorities."

While I agree that the variables chosen are appropriate, I am conflicted about the exclusion of race/ethnicity in the model. This is especially interesting because the developers state that the variables included in the model were included because they are known to have an independent association with differences in scores. The same can be said for race and ethnicity but the developers chose not to include this variable in the model.

Panel Member #5: Reasonable steps taken to develop the risk model. However, the c-stat of the model is poor at 0.58

Panel Member #6: The risk adjustment model includes age, comorbidity score, next-of-kin relationship (between survey respondent and patient), primary diagnosis at death, and mode of survey administration. The inclusion of next-of-kin relationship and mode of survey administration presumably address the differential quality of response, rather than the conceptual risk of a non-excellent response. The discrimination of the model is low, as c = 0.58.

I wonder if length of stay during the last month preceding death is an important factor, if one were to hypothesize that a brief hospitalization preceding death would be associated with increased risk of perception of poor health care.

Panel Member #8: The risk adjustment approach is acceptable.

Panel Member #9: The developers do an acceptable job with adjustment for several aspects of "clinical risk", but the calibration statistics for the adjustment model are not impressive. However, this seems due to the fundamental weakness of associations between available risk factors and the outcome, rather than any flaw in the developers' model development and testing approach. The authors describe several risk factors as "social risk factors" (e.g., comorbidity) when in fact they are not social risk factors. They seem to have a significant effect of race on the measure, but have chosen to not include race, not on the basis of some careful analysis showing that the effect of race is due to quality of care disparities vs. something else, but on the basis of a policy preference in the VA and a brief document of AHRQ guidance. Since biases in response style (e.g., propensity to use an extreme rating, regardless of survey context) are well-known, and are a sound basis for adjusting on the basis of race or ethnicity or culture, the developers can be faulted here for just assuming that any effects of race represent quality of care disparities rather than analyzing available data (including published data by others) to make some determination of whether an effect of race represents a quality disparity (then don't adjust) or a response style difference unrelated to quality of care (then they should adjust).

For cost/resource use measures ONLY: NA – not a cost / resource use measure

17. **Are the specifications in alignment with the stated measure intent?**

    ☐ Yes ☐ Somewhat ☐ No (If "Somewhat" or "No", please explain)

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

VALIDITY: TESTING

19. **Validity testing level: ☐☒☐☒ Measure score ☐☒☐☒ Data element ☒☐ Both**
20. **Method of establishing validity of the measure score:**

    ☐ Face validity

    ☒☒ Empirical validity testing of the measure score

    ☐ N/A (score-level testing not conducted)

21. **Assess the method(s) for establishing validity** The methods used were appropriate – linking the overall rating that represents the proposed measure to other indicators of quality of end-of life care.

    Submission document: Testing attachment, section 2b1.2

    Panel Member #1: No issues**.**

    Panel Member #3: Developers assessed associations between various desirable care processes and patient-level performance scores.

    Panel Member #4: Methods used included: a qualitative analysis of families' written BFS comments and the relationship between comments and BFS items; and discriminant validity (the associations between various facility-level process measures (aka, quality of care indicators or interventions) and facility-level BFS Performance Measure scores. I have no concerns with these methods.

    Panel Member #5: The types of tests seem to be appropriate given the measure.

- discriminant validity… associations between various facility-level process measures (aka, quality of care indicators or interventions) and facility-level BFS Performance Measure
- compared the BFS-PM to individual BFS items included on the survey to evaluate differences in family ratings in specific areas of EOL care versus the BFS-PM rating
- Using patient-level data (N=84,616) and facility-level data (N=146), we ran nine separate logistic/linear regressions adjusted for nonresponse bias and patient case-mix.' [p11-12]

Panel Member #6: Discriminant validity was assessed by comparing hospitalizations with versus without palliative care.

Panel Member #7: Patient and family level associations between the BFS and other related quality measures were assessed using (?) mixed linear model regressions adjusted of non-response bias and case mix.

Panel Member #8: The developer reported a qualitative study that assesses the validity of the global survey item. For measure score validity testing, the developer assessed if the BFS-PM is correlated with other quality indicators such as palliative consolation and emotion support. Both are commonly used approaches and acceptable for validity testing. Again, the description of measure score validity testing could have been more clear, particularly in terms of how logistic and linear regression models were specified.

Panel Member #9: The methods used were appropriate – linking the overall rating that represents the proposed measure to other indicators of quality of end-of life care.

22. **Assess the results(s) for establishing validity** Results of the validity testing showed acceptable levels of relationship between the single-item overall assessment and either survey-based or chart-based indicators of process or structure indicators of quality of care.

Submission document: Testing attachment, section 2b1.3 No issues.**, 2b1.4**

Panel Member #3: Expected associations were present and in the expected directions.

Panel Member #4: "The BFS has consistently shown significant associations (p<0.001) using logistic/linear regression tests with quality of care indicators based on the empirical literature and "Best Practices" as outlined in the National Consensus Project for Quality Palliative Care Clinical Guideline (presence of a palliative consult at death, death in an inpatient hospice unit, chaplain and bereavement contacts with patients and family members)." I have no concerns.

Panel Member #5: It appears, measure score results are presented, but not critical data element results are not presented. Regarding the aforementioned, the results are acceptable.

[measure score?]

- The BFS has consistently shown significant associations (p<0.001) using logistic/linear regression tests with quality of care indicators based on the empirical literature and "Best Practices…"
- Table A below, nonresponse and patient case-mix adjusted patient-level BFS-PM scores are consistently higher when patients receive these process measures/quality indicators….' [p12]
- Table B below, nonresponse and patient case-mix adjusted facility-level BFS-PM scores are consistently higher for when patients receive these quality indicators…' [p13]

Panel Member #6: Receipt of palliative care was associated with higher adjusted odds of patient-level performance measure response of "excellent."

Panel Member #7: All patient and facility level associations with other quality measures were statistically significant, however, some facility level beta coefficients were small (e.g. bereavement contact with family, death in a hospice/palliative care unit, chaplain contact with Veteran or family member). Nevertheless, the pattern is positive and compelling.

Panel Member #8: The developer reported BFS-PM was positively associated with valence ratings for data element validity. Strong association was also reported between other quality indicators and BFS overall rating of EOL care. Similar results were obtained for facility lever scores.

Panel Member #9: Results of the validity testing showed acceptable levels of relationship between the single-item overall assessment and either survey-based or chart-based indicators of process or structure indicators of quality of care.

23. **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)

24. **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**

Panel Member #9: The key phrase here is ALL critical data elements, and the question of whether all of the risk adjustment variables are "key" data elements. The reliability and validity of these data elements was not formally tested. Given their nature and their source, though, there are **not really significant questions about either reliability or validity of the risk-adjustment variables**.

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

Panel Member #5: While the box was checked that data elements testing was performed, I did not locate such test results.

25. **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 **are** threats to validity and/or relevant threats to validity were **not assessed OR** 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 **is required**; if not conducted, should rate as INSUFFICIENT.)

26. **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: Issue of low validity in the prior review. Better odds ratio results provided.

Panel Member #3: I chose insufficient because the description of the case mix weighting adjustment does not make sense to me and I'm hoping the developers can clarify what was done (see comments above). I would likely score the measure as high otherwise.

Panel Member #5: Q22: It appears, measure score results are presented, but not critical data element results are not presented. Regarding the aforementioned, the results are acceptable.

Panel Member #6: The logistic and linear regression models provide good evidence of validity.

Panel Member #7: All patient and facility level associations with other quality measures were statistically significant, however, some facility level beta coefficients were small (e.g. bereavement contact with family, death in a hospice/palliative care unit, chaplain contact with Veteran or family member). Nevertheless, the pattern is positive and compelling.

Panel Member #8: The testing results were in support of the validity of this measure. The developer appropriately incorporated both risk adjustment and response rate weight in determining the final PRO-PM score,

Panel Member #9: . I debated here between moderate and high – I had no question that the measure should pass the validity criterion. My concerns about risk adjustment kept me away from the "high" rating.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **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.**

    Panel Member #9: This measure raises an old question about "endorsement for use". The developers provide reasonable evidence that the measure can be used to compare hospitals as long as a minimum sample size of 56 surveys is reached, but then also state that the measure is intended to be used for quality improvement within hospitals. That intent is given as part of the rationale for not including social risk factors. Since the NQF endorsement is not specific to one particular use, the Standing Committee will have to decide whether the risk adjustment is adequate to allow for fair comparisons among hospitals, in addition to being adequate for internal quality improvement purposes.

Developer Submission

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

BFS_NQF_Measure_Evidence_05.01.2014.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.

No

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1623
Measure Title: Bereaved Family Survey
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: **5/19/2014**

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

☒ Health outcome: **experience with care**
☐ Patient-reported outcome (PRO):
PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors
☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:
☐ Structure:
☐ Other:

____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to* ***1a.3***

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Rating end of life inpatient care as excellent → receiving a palliative care consult, dying in a hospice unit,

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

A growing body of research has underscored the degree to which end-of-life care in the United States needs to be improved. For instance, extensive data from non-VA health care systems clearly demonstrate that physical symptoms like pain, nausea, constipation, and dyspnea are very common. Moreover, clinicians are often unable to recognize these symptoms and manage them adequately. Other studies have found that providers do not communicate with patients about their health care preferences, and that providers' treatment decisions are not always consistent with those preferences.

The challenges of end-of-life care are particularly significant in the U.S. Department of Veterans Affairs (VA) health care system because the VA provides care for an increasingly older population with multiple comorbid conditions. In FY2000, approximately 104,000 enrolled veterans died in the U.S., and approximately 27,200 veterans died in VA facilities. At least 30% of veterans are over age 65 now, and 46% will be over 65 by 2030. Therefore, it is clear that the number of deaths in VA facilities will increase substantially as the World War II and Korean War veterans age. These demographic trends mean that, like other healthcare systems, the VA will face substantial challenges of providing care to veterans near the end of life.

Post-death surveys of family members have emerged as an essential strategy for assessing the quality of end-of-life care in a variety of health care settings. There are at least 5 reasons why post-death family surveys are essential an essential part of an effective measurement strategy. First, family surveys can assess the care of all veterans who die, even those whose prognosis is uncertain, and who therefore might not be identified as "terminally ill" in a prospective assessment. Second, family surveys also avoid challenges of data collection from patients near the end of life, in whom cognitive impairment is common, particularly in ICU deaths. Third, family surveys can retrospectively assess care at the time of death, an emotionally difficult time when data collection from patients or families may be felt to be unacceptably intrusive. Fourth, families' assessments offer an essential source of data because one key outcome—support provided to family members after a patient's death—can only be assessed by after a patient's death. Fifth, surveys of family members provide an essential source of data to assess the support that is provided to family members themselves. For all of these reasons, therefore, post-death assessments by family members will offer an essential source of data that define the quality of end-of-life care that the VA is able to provide for its veterans and their family members.

***Note**: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.*

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. **Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes**. Include all the steps between the measure focus and the health outcome.

1a.3.1. **What is the source of the systematic review of the body of evidence that supports the performance measure?**

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

National Consensus Project for Quality Palliative Care. Clinical practice guidelines for quality palliative care. 2nd ed. Pittsburgh (PA): National Consensus Project for Quality Palliative Care; 2009. 80p.

http://www.guideline.gov/content.aspx?id=14423&search=palliative

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Guideline 7.1 Signs and symptoms of impending death are recognized and communicated in developmentally appropriate language for children and patients with cognitive disabilities with respect to family preferences. Care appropriate for this phase of illness is provided to patient and family.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade: N/A

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

n/a

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*):

n/a

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes *→ complete section 1a.7*

☐ No *→ report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7*

____

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

n/a

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

n/a

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

n/a

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

n/a

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*):

n/a

Complete section 1a.7

____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7 ____________________________

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:
n/a

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.
n/a

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: 2008-2013

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)
1 cohort study

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)
Study flaws - the study is conducted in a VA population, whose demographic characteristics are atypical of the larger U.S. population. Additionally, the study relies on families´ perceptions of care rather than on direct assessments of perceptions. However, retrospective surveys of family members have several advantages over patient assessments. For instance retrospective surveys can assess the care of patients whose prognosis is uncertain and who therefore might not be prospectively identified as ''terminally ill.'' They also make it possible to examine the care of patients who are unable to respond to surveys or questionnaires, which is important because cognitive impairment is present in at least 50% of inpatients in the last weeks of life.27 Retrospective surveys can also provide insights into the care that was delivered at the time of death, when prospective data collection from patients or families may be unacceptably intrusive. These surveys offer the only way to assess the care that is provided to the family after a patient's death.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

n/a

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

n/a

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

n/a

____

1a.8 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.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1b. Performance Gap

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

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

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.*

The measure is used to improve the quality of end-of-life care received by Veterans and family members during the last month of life in a VA inpatient setting.

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.*

The VA has not yet developed and implemented extensive national measures of the quality of end-of-life care it provides to Veterans. There are at least 3 reasons why this is essential. First, a system-wide quality measurement strategy would make it possible to define and compare the quality of end-of-life care at each facility and to identify opportunities for improvement. Second, facilities and VISNs would be able to monitor the effectiveness of efforts to improve care locally and nationally. Third, a system-wide measurement strategy will help the VA to recognize facilities that provide outstanding end-of-life care, so that successful processes and structures of care can be identified and disseminated throughout the VA.

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.

Edes, T., Shreve, S., Casarett, D. Increasing access and quality in Department of Veterans Affairs care at the end-of-life: A lesson in change. Journal of the American Geriatrics Society. 2007;55:1645-49.

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.*

In the testing attachment, the developers provide figures demonstrating significant variation in BFS-PM scores across VA facilities nationwide and over time. The developer also includes citations to a number of published articles that have documented statistically and clinically significant differences in performance with regard to :

1) receipt of a palliative consult;
2) death in an inpatient hospice unit vs. other inpatient venues;
3) diagnoses at the end of life;
4) gender;
5) receipt of aggressive care;
6) chaplain interaction;
7) bereavement support after death;
8) DNR order at time of death;
9) race/ethnicity;
10) family involvement; and
11) timing of palliative consults.

In the validity testing attachment, the developer provides results from 2017 (n=146 facilities) demonstrating a 65% mean overall score, a score range from 13% - 100%, and IQR of 85 and 72.

Disparities

The developer has documented the presence of significant racial/ethnic disparities on the BFS-Performance Measure in previously published work. The developer also shares analysis of Family Assessment of Treatment at the End of life (FATE) results that indicate ethnicity (white vs. nonwhite) and older age were independently associated with higher FATE scores

In a regression model that included adjustments for nonresponse and patient case mix, patients who received a consultation had significantly higher scores than those who did not (63% vs 49%; OR=1.80; 95% CI=1.72-1.88; P<0.001). Race/Ethnicity (non-Hispanic white vs all other race/ethnicity) was also associated with a higher BFS score (62% versus 48%) after adjusting for nonresponse bias and patient case mix.

Higher BFS scores were also seen on the following process measures:

Table A. Adjusted Associations Between Quality Indicators and Patient-level BFS Overall Rating of EOL Care for FY10-FY17, N=48,785

Process Measure/Satisfaction with Quality Indicator	Patient-Level PM Score with (Yes) and without (No) Receipt of Process Measure/Satisfaction with quality indicator (Always)		Odds Ratio	95% Confidence Interval	P-value
	YES/Always	NO/Usually, Sometimes, Never			
Palliative Care Consult prior to death	63	49	1.80	1.72-1.88	<0.001
Death in a Hospice/Palliative Care Unit	68	55	1.73	1.67-1.80	<0.001
Chaplain Contact with Veteran or Family	61	52	1.44	1.35-1.53	<0.001
Bereavement Contact with Family	61	58	1.12	1.07-1.17	<0.001
Staff took time to listen	76	17	15.53	14.69-16.42	<0.001
Staff provided patient with wanted medication/medical treatment	73	18	12.36	11.65-13.11	<0.001

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

Casarett DJ, Pickard AP, Bailey FA, et al. Important aspects of end-of-life care for veterans: Implications for measurement and quality improvement. Journal of Pain & Symptom Management. 2008;35:115-125.

1. Casarett DJ, Pickard AP, Bailey FA, et al. A nationwide VA palliative care quality measure: the family assessment of treatment at the end of life. Journal of Palliative Medicine. 2008;11:68-75.
2. Casarett DJ, Pickard AP, Bailey FA, et al. Do palliative care consultations improve patient outcomes? Journal of the American Geriatric Society. 2008;56:593-99.
3. Alici, Y, Smith D, Lu H, Henderson H, Asch D, Casarett D. Families' Perceptions of Veterans' Distress Due to Post-Traumatic Stress Disorder-Related Symptoms at the End of Life. Journal of Pain and Symptom Management 2010 39(3): 507-514.
4. Lu H, Trancik E, Bailey FA, Ritchie C, Rosenfeld K, Shreve S, Furman C, Smith D, Wolff C, Casarett D. Families Perceptions of End-of-life Care in Veterans Affairs versus non-Veterans Affairs Facilities. Journal of Palliative Medicine. 2010, 13(8): 991-996.

Kutney-Lee A, Carpenter J, Smith D, Thorpe J, Tudose A, Ersek M. (2018). Case-Mix Adjustment of the Bereaved Family Survey. American Journal of Hospice and Palliative Medicine, 1-8. doi: 10.1177/1049909117752669. VAMC facility characteristics vary widely across the nation and those characteristics have an impact on BFS scores. To provide fair comparisons across facilities, the objective of this study was to develop a case-mix adjustment model for the BFS, and to examine changes in facility-level scores following adjustment. Following adjustment using model-based propensity weighting, the mean change in BFS-Performance Measure score across facilities was -0.6 with a range of -2.6 to 0.6. The scores of 108 facilities decreased by less than 1 percentage point, while 55 facilities changed within +0.5 percentage points of their unadjusted score. On average, facilities that benefited most from adjustment cared for patients with greater comorbidity burden and were located in urban areas in the Northwest and Midwestern regions of the country.

Manfredi L, Lorenz K, Smith D, Ersek M, Gale R. (2017). Do family comments received on quality of care surveys help systematically understand performance? Evaluating family comments on the bereaved family survey (BFS). Journal of Pain and Symptom Management, 53(2) 451-452. doi: 10.1016/j.jpainsymman.2016.12.286

Carpenter J, McDarby M, Smith D, Johnson M, Thorpe J, Ersek M. (2017). Associations between timing of palliative care consults and family evaluation of care for Veterans who die in a hospice/palliative care unit. Journal of Palliative Medicine, 20(7), 745-751. doi: 10.1089/jpm.2016.0477 After adjustment for patient and facility characteristics, family members of veterans whose first PCC occurred 91-180 days prior to death were more likely to rate overall care as “excellent” compared with those whose PCC occurred 0-7 days prior to death, 67.9 v. 62.1%, respectively (Adjusted Odds Ratio=1.37; 95% confidence interval (CI) 1.08-1.73). Earlier PCC is associated with greater family satisfaction with care. Strategies aimed at conducting PCC earlier in life limiting illness are needed.

Ersek M, Miller S, Wagner T, Thorpe, J, Smith, D, Levy C, Gidwani R, Faricy-Anderson K, Lorenz K, Kinosian B, Mor V. (2017). Association between aggressive care and bereaved families’ evaluation of end-of-life care for Veterans with non-small cell lung cancer who died in Veterans Affairs facilities. Cancer. doi: 10.1002/cncr.20700.

Kutney-Lee A, Smith D, Thorpe J, del Rosario C, Ibrahim S., Ersek, M. (2017). Race/Ethnicity and End-of-Life Care Among Veterans. Medical Care, 55(4) 342-351. doi: 10.1097/MLR.0000000000000637 Statistically significant differences were observed by race/ethnicity on 1 of the 4 end-of-life quality indicators: black Veterans were less likely than whites to receive a chaplain consult (77% vs. 79%; adjusted OR, 0.83, 95%CI, 0.73-0.94; p=0.004). Among the 15 Bereaved Family Survey items, less favorable outcomes were observed for black, Hispanic, and other racial/ethnic minorities on 12, 8 and 5 items, respectively. In comparison to whites, minority Veterans were less likely to report excellent overall care by the following odds ratios: 0.57 (95%CI, 0.53-0.61; p<0.001) for blacks, 0.85 (95%CI, 0.76-0.94; p=0.002) for Hispanics; and 0.83 (95%CI, 0.71-0.97; p=0.02) for other races/ethnicities. Our study found marked racial/ethnic disparities in the quality of end-of-life care in a national sample of Veterans who receive care in the equal-access VA healthcare system. Family perceptions are a critical component of evaluating equity and quality of care at the end of life.

Thorpe, J. M., Smith, D., Kuzla, N., Scott, L., & Ersek, M. (2016). Does Mode of Survey Administration Matter? Using Measurement Invariance to Validate the Mail and Telephone Versions of the Bereaved Family Survey. Journal of pain and symptom management, 51(3), 546-556. doi: 10.1016/j.jpainsymman.2015.11.006.

Wachterman, M. W., Pilver, C., Smith, D., Ersek, M., Lipsitz, S.R., Keating, N.L. (2016). Quality of end-of-life care provided to patients with different serious illnesses." JAMA Internal Medicine,176(8),1095-1102. doi: 10.1001/jamainternmed.2016.1200. The adjusted proportion of patients receiving a palliative care consult was highest among cancer patients (69.8%) and dementia patients (61.5%), while less than half of patients with ESRD, cardiopulmonary failure, and frailty received such consults (P<.001). The adjusted proportion of patients dying in the ICU was lowest among cancer patients (16.1%) and dementia patients (11.0%), while about one-third of patients with frailty, ESRD, and cardiopulmonary failure died in the ICU (P<.001). The adjusted proportion of patients with a DNR order at death was highest among cancer patients (94.6%) and dementia patients (93.5%), compared to 87-88% for other conditions (P<.001). The adjusted proportion of family members reporting that care in the last month of life was “excellent” was highest among those of patients with cancer (57.7%) and dementia (59.2%) and lowest among patients with ESRD (53.8%) and cardiopulmonary failure (53.5%) (P<.001).

Ersek, M., Thorpe, J., Kim, H., Thomasson, A., Smith, D. (2015). Exploring End-of-Life Care in Veterans Affairs Community Living Centers. Journal of the American Geriatrics Society, 63(4), 644-650. doi: 10.1111/jgs.13348. Family evaluation of overall EOL care and quality of EOL care indicators for Veterans who died in CLCs were better than those of Veterans dying in acute and intensive care units, but were worse than those dying in hospice/palliative care units.

CONCLUSION: Findings indicate that care in the CLC can be enhanced through the integration of palliative care practices. Future research should identify key elements of enhancing EOL care in nursing homes.

Smith, D., Kuzla, N., Thorpe, J., Scott, L., & Ersek, M. (2015). Exploring Nonresponse Bias in the Department of Veterans Affairs´ Bereaved Family Survey. Journal of palliative medicine, 18(10), 858-864. doi: 10.1089/jpm.2015.0050.

Kutney-Lee, A., Brennan, C. W., Meterko, M., & Ersek, M. (2015). Organization of Nursing and Quality of Care for Veterans at the End of Life. Journal of pain and symptom management, 49(3), 570-577. doi: 10.1016/j.jpainsymman.2014.07.002.

Roza, K. A., Lee, E. J., Meyer, D. E., & Goldstein, N. E. (2015). A survey of bereaved family members to assess quality of care on a palliative care unit. Journal of palliative medicine, 18(4), 358-365. doi: 10.1089/jpm.2014.0172.

Sudore, R. L., Casarett, D., Smith, D., Richardson, D. M., & Ersek, M. (2014). Family involvement at the end-of-life and receipt of quality care. Journal of pain and symptom management, 48(6), 1108-1116. doi: 10.1016/j.jpainsymman.2014.04.001.

Ersek, M., Smith, D., Cannuscio, C., Richardson, D. M., & Moore, D. (2013). A nationwide study comparing end-of-life care for men and women veterans. Journal of palliative medicine, 16(7), 734-740. doi: 10.1089/jpm.2012.0537.

Smith, D., Caragian, N., Kazlo, E., Bernstein, J., Richardson, D., & Casarett, D. (2011). Can we make reports of end-of-life care quality more consumer-focused? Results of a nationwide quality measurement program. Journal of palliative medicine, 14(3), 301-307. doi: 10.1089/jpm.2010.0321.

Casarett, D., Johnson, M., Smith, D., & Richardson, D. (2011). The optimal delivery of palliative care: a national comparison of the outcomes of consultation teams vs inpatient units. Archives of Internal Medicine, 171(7), 649-655. doi:10.1001/archinternmed.2011.87.

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):*

Palliative Care and End-of-Life Care

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

Person-and Family-Centered Care

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

Elderly, Populations at Risk : Veterans

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.)*

http://www.cherp.research.va.gov/PROMISE

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)*

No data dictionary **Attachment:**

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**:** Inpatient_ENG_MALE_Survey_Bereavement_VA_5.30.17.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.

Family or other caregiver

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.

Yes

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.

Addition of nonresponse bias and case mix adjustment as well as additional reliability evidence.

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).*

The numerator is comprised of completed surveys (at least 12 of 17 structured items completed), where the global item question has an optimal response. The global item question asks "Overall, how would you rate the care that [Veteran] received in the last month of life" and the possible answer choices are: Excellent, Very good, Good, Fair, or Poor. The optimal response is Excellent.

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).*

Included are those patients included in the denominator with completed surveys (at least 12 of 17 structured items completed) that receive an optimal response on the global item question.

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

The denominator consists of all inpatient deaths for which a survey was completed (at least 12 of 17 structured items completed), excluding:

1) deaths within 24 hours of admission (unless the Veteran had a previous hospitalization in the last month of life);
2) deaths that occur in the Emergency Department (unless the Veteran had a prior hospitalization of at least 24 hours in the last 31 days of life);

Additional exclusion criteria include:

1) Veterans for whom a family member knowledgeable about their care cannot be identified (determined by the family member´s report); or contacted (no current contacts listed or no valid addresses on file);
2) absence of a working telephone available and valid mailing address to the family member.

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).*

The purpose of this measure is to assess families´ perceptions of the quality of care that Veterans received from the VA in the last month of life. The BFS consists of 19 items (17 structured and 2 open-ended). The BFS items were selected from a longer survey that was developed and validated with the support of a VA HSR&D Merit Award and have been approved for use by the Office of Management and Budget.

Seventeen items in the survey have predefined response options and ask family members to rate aspects of the care that the Veteran received from the VA in the last month of life. These items cover areas of care such as communication, emotional and spiritual support. Two additional items are open-ended and give family members the opportunity to provide comments regarding the care the patient received.

A growing body of research has underscored the degree to which end-of-life care in the United States needs to be improved. The challenges of end-of-life care are particularly significant in the U.S. Department of Veterans Affairs Health Care system because he VA provides care for an increasingly older population with multiple comorbid conditions. In FY2000, approximately 104,000 enrolled Veterans died in the U.S., and approximately 27,200 Veterans died in VA facilities. At least 30% of the Veterans are over age 65 now, and 46% will be over 65 by 2030. Therefore, it is clear that the number of deaths in VA facilities will increase substantially as the World War II and Korean War Veterans age. These demographic trends mean that, like other healthcare systems, the VA will face substantial challenges of providing care to Veterans near the end-of-life.

The VA has addressed this challenge aggressively in the last 5 year, however the VA has not yet developed and implemented measures of the quality of end-of-life care it provides to Veterans. There are at least 3 reasons why adoption of a quality measurement tool is essential. First, it would make it possible to define and compare the quality of end-of-life care at each VA facility and to identify opportunities for improvement. Second, facilities and VISNs (geographic service divisions within the VA system) would be able to monitor the effectiveness of efforts to improve care locally and nationally, and would enable monitoring of the impact of the Comprehensive End of Life Care Initiative, ensuring that expenditures are producing improvements in care.

Third, it will help the VA to recognize those facilities that provide outstanding end-of-life care, so that successful processes and structures of care can be identified and disseminated throughout the VA.

The BFS´s 17 close-ended items ask family members to rate aspects of the care that the Veteran received from the VA in the last month of life. These items cover areas of care such as communication, emotional and spiritual support, pain management and personal care needs. Two additional items (not used in scoring) are open-ended and give family members the opportunity to provide comments regarding the care the patient received. The BFS has undergone extensive development and has been pilot-tested for all inpatient deaths in Q4FY2008 in seven VISNs (1,2,4,5,8,11, and 22). As of October 1, 2009, Q1FY2010, all inpatient deaths in all VISNs were included in the project.

The indicator denominator is comprised of the number of Veterans who die in an inpatient VA facility (intensive care, acute care, hospice unit, nursing home care or community living center) for whom a survey is completed. Completed surveys are defined as those with at least 12 of the 17 structured items completed.

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

- Veterans for whom a family member knowledgeable about their care cannot be identified (determined by family member´s report)
- Absence of a current address and/or working telephone number for a family member or emergency contact.
- Deaths within 24 hours of admission without a prior hospitalization of last least 24 hours in the last 31 days of life.
- Deaths that occur in the operating room during an outpatient procedure.
- Deaths due to a suicide or accident
- Surveys in which less than 12 items were answered.

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.)*

Name, address, and phone number of patient´s family member or emergency contact are required for determining exclusion. In addition, information regarding the patient´s admission(s) during the last 31 days of life, and including length of stay are also required to determine exclusion.

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.)*

Variables necessary to stratify the measure are VISN, facility, quarter, year, outcome. VISN refers to "Veterans Integrated Service Network" and is a geographic area of the country where a facility is located. Facility is the actual VA medical center or affiliated community living center where the Veteran died. Quarter is the 3 month time period in which the patient died. Year is the VA fiscal year (runs from Oct 1 to Sept 30). Outcome refers to whether or not a survey was completed.

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.)*

The purpose of this measure is to assess families´ perceptions of the quality of care that Veterans received from the VA in the last month of life. The BFS consists of 19 items (17 structured and 2 open-ended). The BFS items were selected from a longer survey that was developed and validated with the support of a VA HSR&D Merit Award and have been approved for use by the Office of Management and Budget.

Seventeen items in the survey have predefined response options and ask family members to rate aspects of the care that the Veteran received from the VA in the last month of life. These items cover areas of care such as communication, emotional and spiritual support. Two additional items are open-ended and give family members the opportunity to provide comments regarding the care the patient received.

A growing body of research has underscored the degree to which end-of-life care in the United States needs to be improved. The challenges of end-of-life care are particularly significant in the U.S. Department of Veterans Affairs Health Care system because he VA provides care for an increasingly older population with multiple comorbid conditions. In FY2000, approximately 104,000 enrolled Veterans died in the U.S., and approximately 27,200 Veterans died in VA facilities. At least 30% of the Veterans are over age 65 now, and 46% will be over 65 by 2030. Therefore, it is clear that the number of deaths in VA facilities will increase substantially as the World War II and Korean War Veterans age. These demographic trends mean that, like other healthcare systems, the VA will face substantial challenges of providing care to Veterans near the end-of-life.

The VA has addressed this challenge aggressively in the last 5 year, however the VA has not yet developed and implemented measures of the quality of end-of-life care it provides to Veterans. There are at least 3 reasons why adoption of a quality measurement tool is essential. First, it would make it possible to define and compare the quality of end-of-life care at each VA facility and to identify opportunities for improvement. Second, facilities and VISNs (geographic service divisions within the VA system) would be able to monitor the effectiveness of efforts to improve care locally and nationally, and would enable monitoring of the impact of the Comprehensive End of Life Care Initiative, ensuring that expenditures are producing improvements in care. Third, it will help the VA to recognize those facilities that provide outstanding end-of-life care, so that successful processes and structures of care can be identified and disseminated throughout the VA.

The BFS´s 17 close-ended items ask family members to rate aspects of the care that the Veteran received from the VA in the last month of life. These items cover areas of care such as communication, emotional and spiritual support, pain management and personal care needs. Two additional items (not used in scoring) are open-ended and give family members the opportunity to provide comments regarding the care the patient received. The BFS has undergone extensive development and has been pilot-tested for all inpatient deaths in

Q4FY2008 in seven VISNs (1,2,4,5,8,11, and 22). As of October 1, 2009, Q1FY2010, all inpatient deaths in all VISNs were included in the project.

The 17 structured items of the Bereaved Family Survey are scored as either "1" (optimal response) or "0" (all other answer choices). A score of "1" indicates that the family member perceived that the care they and/or the Veteran received was the best possible care (Excellent). A score of "0" reflects all other possible responses (Very good, Good, Fair, Poor). Items are coded as missing if respondents cannot or refuse to answer the item. Thus, the score for each item can be expressed as a fraction corresponding to the number of families who reported that the Veteran received optimal care (numerator), divided by the number of valid, non-missing responses for that item (denominator). Similarly, the score for the 17-item survey is calculated based on the global question item (Overall, how would you rate the care received in the last month of life? - Excellent, Very Good, Good, Fair, Poor). The global item is scored as the # of optimal responses/# of valid, non-missing responses for all completed surveys (12 of 17 structured items answered). This scoring system produces a facility- or VISN-level score that reflects the proportion of Veterans who received the best possible care overall (BFS score) and in specific areas corresponding to BFS items (e.g. pain management, communication, personal care, etc.).

We then add nonresponse and patient case mix weights to the model. All adjusted scores are reported.

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.

All inpatient deaths identified through the VA Electronic Medical Records and provided at the VISN-level. Inpatient deaths are screened for inclusion/exclusion criteria (items 2a.9). Approximately 4 weeks after death, an introductory letter and survey is sent to all the identified Next of Kin (NOK) of all eligible Veterans with a toll-free opt-out line. Approximately 6 weeks after death, a follow-up postcard is sent to encourage participants to complete the survey. Approximately 10 weeks after death, an attempt by telephone to the identified NOK for all non-respondents is made to confirm receipt of the survey and the option to complete the survey by telephone is provided.

NOKs are identified in the following order:

1) Official listing of Next of Kin (NOK)
2) Primary decision-maker as documented in a Social Work or other clinical note
3) Primary decision-maker as documented in an Advance Directive note (this would take precedence over #2 if the Advance Directive note is more recent)
4) Designated Durable Power of Attorney for Health Care

The survey may be administered during a contact telephone call, or at a later time, depending on the family member's preference and availability. Previous experience has shown that some family members will prefer to do the interview immediately, whereas others may prefer a different time or may simply want more time to collect their thoughts.

The survey should be administered in a single telephone call (introduction, consent, survey questions, and conclusion). Interviews are best done in a quiet place, with a minimum of background noise and freedom from interruptions. Wherever possible, doors should be closed and pagers or cell phones should be set to vibrate mode to avoid unnecessary distractions.

The following guidelines should be used when conducting the interviews:

nd carefully. This is particularly important when the family member is an older adult, or when the family member is using a

stop and respond to requests for clarification of questions.

our own questions.” When a family member is confused by a question, do not give your explanation of what the question his is a very natural reaction, it means that you may influence the family’s responses in subtle but important ways. You can stion and gently ask that the family member answer to the best of his or her ability.

r families to exhibit signs of distress (e.g., sadness, tearfulness) and even anger. This is a normal part of the interviewing ny people, who still usually find the interview to be valuable.

Additional demographic data is collected using the VA´s Electronic Record System.

A minimum sample size of 30 respondents is suggested to make comparisons between groups (facilities, VISNs).

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.

Surveys are sent to the listed Next of Kin of in the medical records approximately 4-6 weeks after death of the Veteran. NOKs are provided with an introductory letter and a toll-free line if they wish to either opt-out or complete the survey over the telephone.

There is no minimum response rate.

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

If other, please describe in S.18.

Instrument-Based 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.

For 2a1.25 - Family reported data/survey.

For 2a1.26 - Bereaved Family Survey

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 in attached appendix at A.1

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

Facility, Other

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

Inpatient/Hospital, Post-Acute 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.)*

n/a

2. Validity – See attached Measure Testing Submission Form

BFS_TestingAttachment_052020.docx

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*)**:** 1623
Measure Title: Bereaved Family Survey
Date of Submission: **5/20/2020**
Type of Measure:

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

*cell intentionally left blank

2. 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: Bereaved Family Survey results	☒ other: Bereaved Family Survey results

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*).

This is a primary data set consisting of the VA's Bereaved Family Survey and linked electronic health record data. The denominator is the number of completed Bereaved Family Surveys (BFS) The numerator is the number of family members who answered "excellent" on the BFS Performance Measure (PM) item (Q.18). Please see 1.5 below for specifics.

1.3. What are the dates of the data used in testing? October 2009-February 2019

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
x other: individual patient	☒ other: individual patient

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*)

84,616 completed Bereaved Family Surveys out of 170,764 inpatient deaths were used for testing and analysis at the patient-level; 146 facilities were used for testing at the facility-level on the Bereaved Family Survey (BFS). Facility descriptors (complexity level, urban/rural setting) and patient demographics (age, diagnoses, race/ethnicity, etc.) were also used for testing and analysis. All Veterans who die in one of all 146 Veterans Affairs Medical Centers (VAMCs) nationally were included in the sample, excluding 1) deaths within 24 hours of admission (unless the Veteran had a previous hospitalization of at least 24 hours in the last month of life); 2) Veterans for whom a next of kin (NOK) is not knowledgeable about the care received during the last month of life; 3) Veterans for whom a NOK is not listed in the medical record (N=682; 0.4%); 4) Veterans for whom a NOK has incomplete or incorrect contact information (N=15,174; 8.9%); and 5) Veterans for whom a NOK did not complete at least 12 of 17 structured BFS items (N=1,340; 0.8%). This last determination is made a-priori and is intended to limit the amount of missing data. Extensive data imputation methods for missing data can jeopardize observing true associations between variables of interest (Brick & Kalton, *Statistical Methods in Medical Research*, 1996).

VAMCs are distributed across the nation, including Alaska, Hawaii and Puerto Rico with 90% of facilities in urban areas. The number of VAMC deaths per year also vary (mean number of deaths per year=119; range 1-610) as well as facility complexity (82% highly complex facilities). The VAMC complexity level is a VA administrative categorization based on a weighted combination of seven factors which include patient volume and risk, extent of teaching and training activities, available clinical services and amount of research involvement. In the VA healthcare system, "low complexity" facilities are generally comprised of community living centers (i.e. VA nursing homes) that may have inpatient hospice units. These settings generally have the highest response rates to the survey and are also more likely to have higher scores on the BFS-PM. To account for these differences when making facility-level comparisons, we stratify facility-level scores by facility complexity level prior to reporting.

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*)

84,616 completed Bereaved Family Surveys out of 170,764 inpatient deaths were used for testing and analysis at the patient-level, facility-level, and national-level on the Bereaved Family Survey (BFS). Facility descriptors (complexity level, urban/rural setting) and patient demographics (age, diagnoses, race/ethnicity, etc.) were also used for all testing levels. All Veterans who die in one of 146 Veterans Affairs Medical Centers nationally were included in the sample, excluding 1) deaths within 24 hours of admission (unless the Veteran had a previous hospitalization of at least 24 hours in the last month of life); 2) Veterans for whom a next of kin (NOK) is not knowledgeable about the care received during the last month of life; 3) Veterans for whom a NOK is not listed in the medical record; 4) Veterans for whom a NOK has incomplete or incorrect contact information; and 5) Veterans for whom a NOK did not complete at least 12 of 17 structured BFS items. The majority of Veteran deaths were in the Southeast region of the US including Puerto Rico (11.3% of all deaths) and smallest percent of Veteran deaths occurred in VAMCs in the Northwest region of the US including Alaska (2.9% of all deaths). Veterans where mostly non-Hispanic white (69.6%) males (97.4%) with a mean age of 75 years (range in years 22-110).

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.

There may be differences in those who respond and do not respond to the BFS. Nonresponse bias testing included all inpatient deaths, regardless of BFS completion (N=170,764). Other forms of testing only included Veterans for whose NOK completed the BFS(N=84,616) (validity testing, exclusions, and risk adjustment). A random sample of 93 patients was used to test test-retest reliability at the patient level of the BFS itself.

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.

Data collection from the electronic medical record (EMR) included the following social risk factors: 1) Veteran's age at the time of death; (2) number of medical comorbidities present at the time of death; (3) Veteran's primary diagnosis on last admission;(4) relationship of Veteran's next-of-kin (i.e., spouse), and (5) model of administration mode (i.e., mail).

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*)

Test-retest reliability as well as internal consistency reliability was used for testing reliability.

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

The performance measure (PM) that we are presenting to NQF for endorsement is the ***Bereaved Family Survey (BFS) Overall Rating of End-Of-Life (EOL) Care***. The PM is item #18 from the 20-item BFS (English and Spanish versions). The total BFS includes 18-forced choice items plus 2 open-ended questions. The BFS Overall Rating of EOL Care item is as follows: *"Overall, how would you rate the care that [the Veteran decedent] received in the last month of life?"* Response options are: Excellent—Very Good—Good—Fair—Poor. The reported PM is calculated as the number of respondents who choose "Excellent" (v. all other responses) divided by the number of completed BFS [defined as surveys with a valid response for item #18 **plus** at least 12 more valid responses on the forced-choice items). Thus, the range of possible facility or national scores is 0 to 100%. We use the percentage of "excellent" overall ratings for reporting to VA leadership and clinicians, and follow "top-box" scoring similar to CAHPS-Hospice surveys and to alleviate any ceiling effects.

The national BFS Overall Rating of EOL Care is calculated on a quarterly and annual basis for benchmarking purposes. However, the facility-level scores (reported on a quarterly and annual basis) are commonly the target for quality improvement efforts.

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*)

I. **Data Element Reliability Testing**

BFS Test-Retest Reliability: To evaluate test-retest reliability as a measure of repeatability (aka, stability) of BFS overall rating over time, we randomly selected 93 BFS respondents who agreed to complete the BFS on a 2nd occasion (30 days apart).

Test-Retest Reliability of Absolute Agreement, Statistical Analysis #1 (Cohen's kappa): For binary outcomes such as the overall BFS score, Cohen's Kappa is a commonly-used measure of absolute agreement between scores measured by multiple raters, or scores measured by stability test-retest agreement within a single respondent measured over multiple occasions. Or alternatively, Cohen's kapa assesses the ability of an item to produce the **same** score on multiple occasions. Cohen's kappa measures the within-respondent error in test-retest repeatability of the overall BFS score.

Test-Retest Reliability of Absolute Agreement, Statistical Analysis #2: two-way RANDOM effects, absolute agreement, single rater/measurement. The Intraclass correlation coefficients (ICC) is a desirable measure of test-retest reliability because it reflects both the degree of correlations and absolute agreement between time1 and time2 measurement occasions. While ICCs were developed for test-retesting of continuous outcomes, recent research suggests ICCs on binary outcomes reasonably approximates traditional, linear, ICC estimates. We chose ICC measures of absolute agreement over Cohen's kappa because the kappa coefficient may yield downward-biased reliability estimates depending the prevalence and variability of responses in the sample on the low values when the ratings suggest high reproducibility ("kappa paradox") (D. V. Cicchetti and A. R. Feinstein, High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 1990; 43: 551–558.). We examined two related forms of ICC test-retest absolute agreement: two-way random intercept ICC (2,1); and two-way mixed effect ICC(3,1). ICC(2,1) assumes BFS respondents were randomly selected from the larger population of respondents with similar characteristic. Both ICC(2,1) and ICC(3,1) described below answer the question: How reliable is the BFS item if measured only one time; as per usual protocol?

Test-Retest Reliability, Statistical Analysis #3, Logistic Regression: Strength of association between time 1 and time 2 BFS overall scores was evaluated using a logistic regression of the following form:

The exponentiation of

$$\beta_1$$

provides an estimate of the odds ratio. The odds ratio may be interpreted as follows: Compared to those who reported BFS=0 at time 1, respondents who reported BFS=1 at time 1 had X times the odds of reporting BFS=1 at time 2.

Test-Retest Reliability, Statistical Analysis #4, Cohen's *d* Effect Size of a 2x2 contingency table: Cohen's d is an effect size indicator based on the standardized different between two means or proportions. In this report, Cohen's d assesses the size of the effect of reporting BFS=1 at time 1 (vs BFS=0) on the probability of reporting BFS=1 at time 2. Cohen's d is the number of standard deviations in the probability of reporting BFS=1 at time 2 for those responding BFS=1 at time 1 compared to the mean probability of reporting BFS=1 at time 2 for those responding BFS=0 at time 1. Cohen suggested that d=0.2 be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size.

Data element reliability testing (Internal Consistency Reliability):
Reliability testing of different modes (telephone versus mail) of Bereaved Family Survey administration was conducted to assess the coherence of results across survey modes (internal consistency reliability). The responses from all responders who completed a mail and telephone survey were used. Measures of homogeneity (e.g., Cronbach) were computed (mathematically equivalent to the average of all possible split-half estimates) as well as exploratory factor analysis.

Facility-level Reliability Testing I (Intraclass Correlation Coefficient): To further establish variability in facility-level BFS scores, we decomposed the within- and between-facility variance in overall BFS score using a mixed-effects logistic regression model. The Intraclass Correlation Coefficient (ICC1) is a signal-to-noise ratio of the between-facility variability relative to the total variability in BFS scores.

Facility-level Reliability Testing II (Spearman-Brown Split-Half Reliability): The reliability of facility-averaged scores was assessed via split-half reliability at the facility-level using the Spearman-Brown prophecy formula. Each facility was randomly split in half and compared to the other half of completed surveys. The Spearman-Brown formula indicates a minimum facility-level sample size of 56 respondents is required to achieve the recommended reliability threshold of 0.70. Over 97% of facilities have sufficient sample size to achieve 70% reliability. Sample size is only being used to show that our sample meets the recommended minimum to achieve the reliability threshold. However, we do not require a sample size minimum when reporting scores.

$$r_{\mathrm{SB}} = \frac{2r_{hh}}{1+r_{hh}}$$

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*)

RESULTS

Data element testing: Test-Retest Reliability

Descriptive Statistics: Overall, 80.4% of the 92 respondents reported a BFS overall score of "excellent" at time 1, and 83.7% reported "excellent" at time 2. There was 85.9% agreement in BFS overall scores of at time 1 and time 2 assessments; 75% agreement at BFS=1 (excellent), and 11% agreement at BFS=0 (p, f, g, vg). 14.1% had discordant responses at time 1 and time 2.

Time 1: BFS Overall Score	Time 2: BFS Overall Score[1] , n, (% of total): 0	Time 2: BFS Overall Score[1] , n, (% of total): 1	TOTAL
0	10 (10.9%)	5 (5.4%)	15 (16.3%)
1	8 (8.7%)	69 (75.0%)	77 (83.7%)
TOTAL	18 (19.6%)	74 (80.4)	92 (100%)

[1] BFS Overall Score: 0 = poor, fair, good, very good. 1 = excellent

Test-Retest Reliability Measures: All measures of test-retest reliability are reported in Table XX below

1. Cohen's kappa for the 92 respondents measured on two occasions is 0.5. According to Cohen's original article, a kappa of 0.5 indicates "moderate" agreement between time 1 and time 2 BFS assessments.
2. ICC(2,1): The 2-way random-effects, absolute agreement, model estimated an ICC(2,1) = 0.52. This is considered a "moderate" level of agreement according to Cohen's original article.

3. Odds Ratio estimate: Compared to those who reported BFS=0 at time 1, respondents who reported BFS=1 at time 1 had 17.2 the odds of reporting BFS=1 at time 2. An odds ratio of this magnitude represents a very strong association between BFS overall scores at time1 and time2.
4. A Cohen's d of 0.8 or greater is considered a 'large' effect size. Our reported Cohen's d of 1.57, therefore, represents a very strong effect of BFS overall score at time 1 on BFS overall score at time 2.

Table XXX. Measures of test-retest reliability between BFS item scores from time 1 and time 2.

Measure	Measures of Test-Retest Reliability	"Rule of thumb" Interpretation
1. Cohens kappa[1]	0.498 (0.26 - .71)	Moderate agreement
2. ICC(2,1)[2]	0.52 (0.36 – 0.66)	Moderate agreement
3. Odds Ratio Estimate[3]	OR = 17.3 (95% CI: 4.7 – 63.3)	Very strong association
4. Cohen's d[4]	1.57	>= 0.8: 'large' effect size

[1] Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

[2] Two-way random-effect Intraclass Correlation Coefficient (ICC) is a measurement of absolute agreement between test-retest of BFS item scores. ICCs from 0.4 – 0.6 are considered "moderate" agreement.

[3] The odds ratio represents the odds of reporting BFS=1 at **time 2** based on BFS overall score at **time 1**. Or: How well does time 1 BFS score predict time 2 BFS score?

[4] The Cohen's d of 1.57 is interpreted as follows: The probability of reporting BFS="excellent" at time 2 is 1.57 standard deviations higher in respondents reporting BFS="excellent" at time 1.
Cohen suggested that d=0.2 be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size.

Summary: Test-Retest Reliability Measures. Taken together, our evaluation demonstrates the test-retest repeatability of BFS scores over time. Overall agreement between time1 and time2 scores was nearly 86%. Both Cohen's kappa and ICC measures of absolute agreement of scores over time demonstrate moderate reliability. This is corroborated by the very large odds ratio (17.2); measuring the strength of association between time 1 and time 2 BFS scores. Further corroboration comes from a Cohen's d of 1.57; an estimate that far exceeds Cohen's d of 0.8 which is considered a "large" effect size.

Data element reliability testing:

Testing was performed on 40,279 family members who completed a BFS telephone survey between October 2009 and September 2012 and 42,378 family members who completed the BFS mail survey between October 2012 and September 2017. The BFS global item dichotomous score (range 0-1), is approximately normally distributed (mean telephone survey score between October 2009 and September 2012: 58, standard deviation 5; mean mail survey score between October 2012 and September 2013: 63, standard deviation 5). Only 2 of 146 VAMCs had a score ≥90, indicating no ceiling effects. Cronbach's alpha for the telephone survey between October 2009 and September 2012 was 0.81 and 0.83 for the mail survey between October 2012 and

September 2017, indicating good homogeneity that is sufficient for between-group comparisons (e.g. comparisons among facilities).

Facility-level Reliability Testing I (Intraclass Correlation Coefficient):

Our analysis demonstrated significant facility-level variation in the latent facility-level scores both for FY10-FY12 -- years during which the BFS was administered predominantly as a phone survey -- (facility-level variance estimate =.15; 95% CI .12-.20; p<0.001) and for FY13-FY17 – when the BFS transitioned to a predominantly mail survey (facility-level variance estimate =.13; 95% CI .09-.20; p<0.001). The ICC1 estimates corresponding to these periods are .04 (95% CI: .03-.06) and .04 (95% CI: .03-.06) respectively. These analyses demonstrate significant facility-level variability in latent facility-level BFS scores.

Facility-level Reliability Testing II (Spearman-Brown Split-Half Reliability):

The estimated SBPH reliability of aggregated facility-level mean scores of 0.89 exceeds the minimum recommended reliability threshold of 0.70 (LeBreton JM; Senter JL. Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 2008;*11*, 815-852). Additionally, based on our estimated ICC1 of 0.04 (4% of variance), the Spearman-Brown formula indicates a minimum facility-level sample size of 56 respondents is required to achieve the recommended reliability threshold of 0.70. Over 97% of facilities have sufficient sample size to achieve 70% reliability.

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?*)

Results from the reliability testing are in accordance with accepted industry standards.

____

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.

Manfredi L, Lorenz K, Smith D, Ersek M, Gale R. (2017). Do family comments received on quality of care surveys help systematically understand performance? Evaluating family comments on the bereaved family survey (BFS). Journal of Pain and Symptom Management, 53(2) 451-452. doi: 10.1016/j.jpainsymman.2016.12.286. To conduct a qualitative analysis of families' written BFS comments and explore the relationship between comments and BFS items, we analyzed a random 5% of written responses (n=341) to the question: "*Is there anything else that you would like to share about the Veteran's care during the last month of life?*" Using codes derived from the BFS, analysts identified which quantitative BFS item(s) responses addressed as well as comment valence (1=positive; 0=neutral; -1=negative). Compound responses (comments relating to more than one BFS item) were disaggregated. Mean/ number of independent comments and average valence were calculated and assessed for correlation with the BFS Performance Measure (PM; overall rating of care). Average number of unique comments to the question was 1.5±0.8 (range: 1 to 6) and average valence was 0.5±1.5 (range: -5 to +6). Comments most frequently related to whether staff was kind/caring/respectful (28.4%) and whether family members were kept informed about the Veteran's condition/treatment (13.5%). Average valence of unique comments ranged from -0.06±0.31 (was Veteran provided with medication/treatment they/you wanted) to 0.21±0.49 (kind/caring/respectful) indicating that although staff are caring, additional communication about treatment preferences may be warranted. BFS-PM was positively correlated with valence ratings (Spearman=0.51; p<0.001).

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)*

The most important test of the survey's usefulness is its discriminant validity for family reported care quality in those with and without palliative care involvement. That is, its ability to distinguish among groups that should, in theory, have different scores. Table 2 below summarizes the extensive analyses that we conducted to examine the associations between various facility-level process measures (aka, quality of care indicators or interventions) and facility-level BFS Performance Measure scores. We have hypothesized that receipt of each of these quality indicators or "Best Practices" should result in a statistically significant higher BFS score.

We developed several chart-derived process variables based on the empirical literature and "Best Practices" as outlined in the National Consensus Project for Quality Palliative Care Clinical Guideline. These variables included: 1) Receipt of a comprehensive palliative consult in the patient's last 90 days of life; 2) Patient or family contact with a chaplain in the last month of life; 3) death in a dedicated inpatient hospice unit; and 4) evidence of emotional support given to a family member up to two weeks post-Veteran death [bereavement contact]. All process variables were dichotomized to reflect the proportion of patients who received each indicator.

We also compared the BFS-PM to individual BFS items included on the survey to evaluate differences in family ratings in specific areas of EOL care versus the BFS-PM rating.

Using patient-level data (N=84,616) and facility-level data (N=146), we ran nine separate logistic/linear regressions adjusted for nonresponse bias and patient case-mix. The independent variables were the process measures and individual BFS item and facility and patient -level BFS % "excellent" was the outcome variable.

Validation of process measure data collection: Currently, all variables are extracted directly from the VA Corporate Data Warehouse (an integrated system of national databases including clinical, administrative and financial data) using standardized algorithms. Prior to 2013, the variables were collected by hand via extensive chart reviews. In the earlier years, all data abstracted from the electronic medical record were collected by trained staff using standardized protocols. Each staff member was required to meet a minimal level of agreement and accuracy prior to collecting data independently, and supervisors conducted regular data quality audits. Depending on the year, nine to twenty staff reviewed an average of 5,000 medical charts annually. Two quality assurance managers checked a 10% random sample of all chart extractions each quarter, and the error rate was < 3%.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
The BFS has consistently shown significant associations (p<0.001) using logistic/linear regression tests with quality of care indicators based on the empirical literature and "Best Practices" as outlined in the National Consensus Project for Quality Palliative Care Clinical Guideline (presence of a palliative consult at death, death in an inpatient hospice unit, chaplain and bereavement contacts with patients and family members).

As can be seen in Table A below, nonresponse and patient case-mix adjusted patient-level BFS-PM scores are consistently higher when patients receive these process measures/quality indicators. Weighted logistic regression analyses demonstrate statistically significant, positive associations between receipt of a quality indicator and patient-level BFS Performance Measure scores.

Table A. Adjusted Associations Between Quality Indicators and Patient-level BFS Overall Rating of EOL Care for FY10-FY17, N=84,616

Process Measure/Satisfaction with Quality Indicator	Patient-Level PM Score with (Yes) and without (No) Receipt of Process Measure/Satisfaction with quality indicator (Always)	Patient-Level PM Score with (Yes) and without (No) Receipt of Process Measure/Satisfaction with quality indicator (Always)	Odds Ratio	95% Confidence Interval	P-value
*	YES/Always	NO/Usually, Sometimes, Never	*	*	*
Palliative Care Consult prior to death	63	49	1.80	1.72-1.88	<0.001
Death in a Hospice/Palliative Care Unit	68	55	1.73	1.67-1.80	<0.001
Chaplain Contact with Veteran or Family	61	52	1.44	1.35-1.53	<0.001
Bereavement Contact with Family	61	58	1.12	1.07-1.17	<0.001
Staff took time to listen	76	17	15.53	14.69-16.42	<0.001
Staff provided patient with wanted medication/medical treatment	73	18	12.36	11.65-13.11	<0.001
Staff were kind, caring and respectful	71	11	19.16	17.76-20.67	<0.001
Staff provided enough emotional support prior to death	80	26	11.27	10.76-11.80	<0.001
Staff provided enough emotional support after to death	75	27	8.09	7.72-8.48	<0.001

*cell intentionally left blank

As can be seen in Table B below, nonresponse and patient case-mix adjusted facility-level BFS-PM scores are consistently higher for when patients receive these quality indicators. The facility-level score is the proportion of family members of deceased Veterans that rated overall end-of-life care as “Excellent”. All 146 facility scores are then averaged into one large national mean, weighted by the number of completed surveys in each facility. Coefficients appear lower when examining facility-level associations due to the loss of variation that occurs with aggregation. Weighted linear regression analyses demonstrate statistically significant, positive associations between receipt of a quality indicator and facility-level BFS Performance Measure scores.

Table B Weighted, Adjusted Associations Between Quality Indicators and facility-level BFS Overall Rating of EOL Care for FY10-FY17, N=146

Process Measure/Satisfaction with Quality Indicator	Process Measure National Facility level %	Process Measure National Facility level SD	β coefficient	95% Confidence Interval	P-value
Palliative Care Consult prior to death	69	12	0.45	0.45-0.46	<0.001
Death in a Hospice/Palliative Care Unit	38	24	0.13	0.12-0.13	<0.001
Chaplain Contact with Veteran or Family	54	11	0.18	0.17-0.19	<0.001
Bereavement Contact with Family	44	17	0.02	0.02-0.03	<0.001
Staff took time to listen	75	5	1.41	1.41-1.42	<0.001
Staff provided patient with wanted medication/medical treatment	80	5	1.57	1.56-1.58	<0.001
Staff were kind, caring and respectful	84	4	1.78	1.76-1.77	<0.001
Staff provided enough emotional support prior to death	65	6	1.21	1.20-1.21	<0.001
Staff provided enough emotional support after to death	70	6	1.17	1.16-1.18	<0.001

*adjusted national facility-level Overall Score mean =59% (SD=8.6)

We have also documented the association between these process measures/ interventions and BFS scores in the following peer-reviewed publications:

1. *Casarett D, Pickard A, Bailey FA, et al. Do palliative consultations improve patient outcomes? J Am Geriatr Soc. Apr 2008;56(4):593-599.* (From Abstract: Interviews were completed with 524 respondents. In a multivariable linear regression model, after adjusting for the likelihood of receiving a palliative consultation (propensity score), palliative care patients had higher overall scores: 65 (95% CI: 62–66) versus 54 (95% CI: 51–56; P < 0.001).
2. *Finlay E, Shreve S, Casarett D. Nationwide veterans affairs quality measure for cancer: the family assessment of treatment at end of life. Journal of clinical oncology. 2008;26(23):3838-3844.* (Receipt of palliative care consult and hospice referral were significantly associated with higher overall scores).
3. *Smith D, Caragian N, Kazlo E, Bernstein J, Richardson D, Casarett D. Can we make reports of end-of-life care quality more consumer-focused? results of a nationwide quality measurement program. J Palliat Med. Mar 2011;14(3):301-307.* (Receipt of palliative care consult, care in a

hospice unit, chaplain contact, emotional support given to a family member post-Veteran death all were significantly associated with higher overall scores).

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)
The BFS has shown, through peer-reviewed analyses, its ability to distinguish among groups that should, in theory, have different scores. The BFS meets the standards of discriminant validity. Finally, palliative care involvement at the end of life improves next of kin reported quality of care delivered.

____

2b2. EXCLUSIONS ANALYSIS

☐ **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*)
All Veterans who die in one of all 146 Veterans Affairs Medical Centers (VAMCs) nationally were included in the sample, excluding 1) deaths within 24 hours of admission (unless the Veteran had a previous hospitalization of at least 24 hours in the last month of life) (N=3,099 or 4%); 2) Veterans for whom a next of kin (NOK) is not knowledgeable about the care received during the last month of life (N=174 or <1%); 3) Veterans for whom a NOK is not listed in the medical record (N=313 or <1%); 4) Veterans for whom a NOK has incomplete or incorrect contact information (N=7,678 or 11%); and 5) Veterans for whom a NOK did not complete at least 12 of 17 structured BFS items (N=646 or <1%).

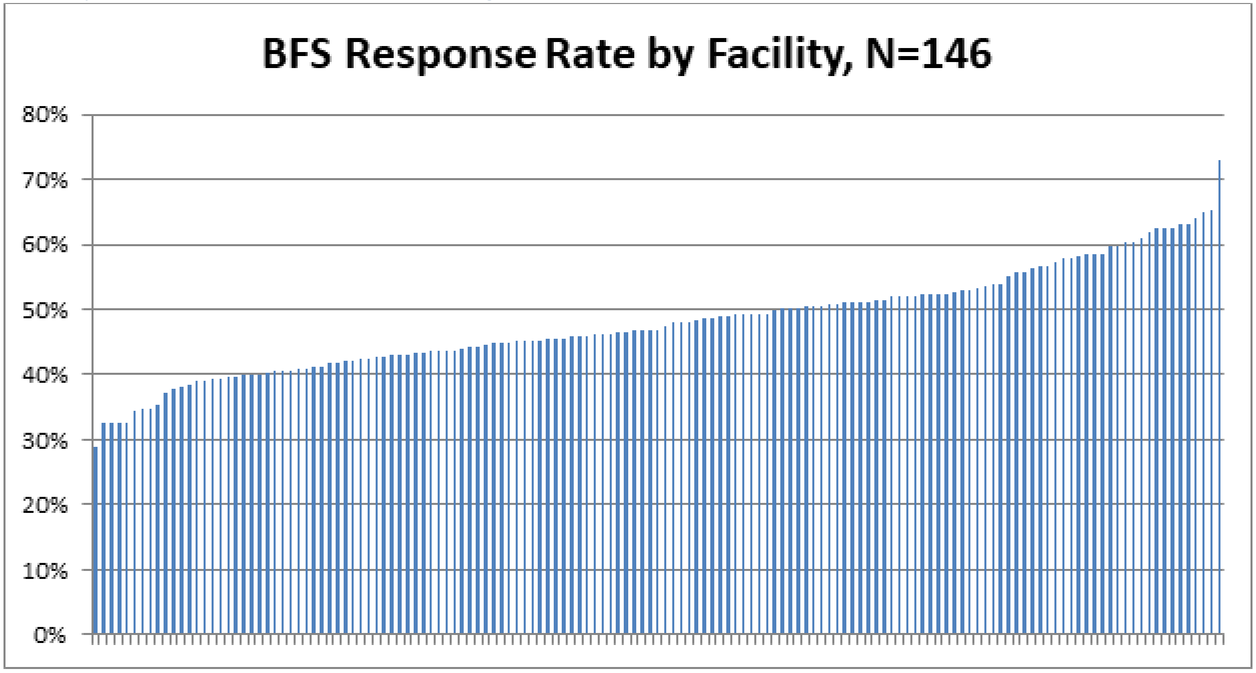
A total of 11,910 Veterans were excluded from the analysis (16%). Veterans who were an inpatient <24 hours before death (4%) were excluded from all analysis including nonresponse bias because they were not in an inpatient setting long enough to have reliable survey data (BFS) or any data at all (quality care indicators/process measures). All remaining exclusions were included in nonresponse bias testing.

To evaluate the effect of nonresponse on BFS-PM scores, associations were examined between several demographic and clinical characteristics and likelihood of survey response. All variables were considered for inclusion in a multivariable model in which the dependent variable was survey completion. This and all subsequent models used robust jackknife standard errors, clustered according to facility (146 clusters). After creating a model to predict the likelihood of response based on patient and clinical characteristics, we applied inverse probability weights to examine their effect on national and facility-level scores.

The survey nonresponse weight used in the model is calculated using the following variables: post-death bereavement contact with a family member, receipt of chaplain contact, receipt of a palliative care consult, death in community living center (i.e. a VA-operated nursing home) or hospice unit, next-of-kin (BFS respondent), Veteran race/ethnicity, Veteran age, and medical comorbidities (obesity, HIV, congestive heart failure, coagulopathy, neurological disorders, and paralysis). The c-statistic obtained during original model development approached 0.65 with an AIC of 24154.07. Scores are weighted following the procedure outlined above. Weights are re-calculated quarterly.

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*)

Rates of response by facility varied, with a mean response rate of 48% (*n*=146) and a range of 29% to 73%. Response rates for facilities with distinct characteristics were also wide-ranging. For example, facilities with an inpatient hospice unit had a higher rate of response when compared to facilities that did not have an inpatient hospice unit (mean response=49% (n=58); range=29%-69% vs. mean response=46% (n=88); range=29%-79%). Also, family members of patients who died in a facility that is categorized as "low complexity" (e.g., non-tertiary care facilities with a majority of non-acute care beds) were more likely to respond to the BFS than family members of patients who died in a "high complexity" facility (mean response=53% (n=60); range=29%-69% vs. mean response=44% (n=86); range=29%-79%). The mean change for all facilities before and after weighting for nonresponse was -2% points, with a range of -10 to +11. Of the 146 facilities in the sample, the scores of 45 facilities (31%) changed more than 2% points in either direction. Twenty-one facilities (14%) did not change their BFS-PM scores.



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*)

The mean change in facility BFS-PM scores after weighting was −2%, (range: −10% to+11%). The scores of 31% of facilities changed more than±2%. The number of facilities meeting hypothetical benchmarks of 60%, 70%, and 80% also changed as a result of weighting for nonresponse. The results underscore the importance of appropriately addressing nonresponse in the use of quality-of-care metrics based on Bereaved Family Survey (BFS) data.

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 5 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.

Prior to reporting of facility-level scores, the BFS-Performance Measure is adjusted for patient case-mix and survey nonresponse, and are stratified by facility complexity level. We examined 5 variables as potential case-mix adjustors: Veteran's age at the time of death (in years), number of medical comorbidities present at the time of death as defined by Van Walraven and colleagues' modification of the Elixhauser score, Veteran's primary diagnosis on last admission (classified into 1 of 15 clinical categories using the Agency for Healthcare Research and Quality Clinical Classification Software), relationship of Veteran's next-of-kin (i.e., spouse), and BFS administration mode (i.e., mail). Additional variables were considered and were either not available (primary language spoken in the home) or not used by other established survey measures (SHEP, CAHPS). Comorbidities and diagnoses are included in the case-mix adjustment model because these health factors were found to have large and statistically significant effects on BFS-PM scores in our prior work (Kutney-Lee et al, 2018, ***American Journal of Hospice & Palliative Medicine***). In this paper, that describes the development of the case-mix model for the BFS-PM, we also found significant differences in facility rankings before and after adjustment for the comorbidity burden of a facility's patients. Further, we include these variables to align with the Center for Medicare and Medicaid Services' (CMS) general approach for adjustment of the CAHPS-Hospice survey (https://www.hospicecahpssurvey.org/globalassets/hospice-cahps/scoring-and-analysis/8-29-2019-updates/cma_public_document-for-website-2018q4-final.pdf), and also in following the recommendation of AHRQ to adjust for patient severity of illness when making facility-level quality comparisons (https://www.ahrq.gov/talkingquality/translate/scores/adjustment-scoring.html).
All reported CAHPS-Hospice survey risk-adjustment variables that were available in VA databases were considered for the BFS case-mix adjustment model. Additional variables were considered and were either not available to us (e.g., primary language spoken in the home, caregiver age, caregiver education) or not used by other established survey measures (CAHPS-Hospice). In order to be considered for the case-mix model, variables must affect how the respondents rate the experience of their care, independent of providers (Elliot MN, Zaslavsky AM, Goldstein E, et al. Effects of survey mode, patient mix and nonresponse on CAHPS Hospital Survey scores. Health Services Research Journal. 2009;44(2 pt 1):501-518).
VA data is pulled from the patient's electronic medical record which has <1% missing data. Decedent age is determined by government record as well as patient report. Patient diagnoses and comorbidities are determined by ICD9/ICD10 codes recorded in the medical record. Next of kin relationship is taken from the patient medical record and determined by both patient and next of kin. Mode of survey completion is recorded by research staff and/or survey administration vendor as soon as the BFS is completed by either mail, telephone or website. All adjustment variables are considered to have high integrity by industry standards and are used widely by VA investigators and leadership.

To examine relationships between case-mix variables and the outcomes of interest, we constructed a set of regression models using logistic regression for the BFS-PM score. Models were fit using raw coefficients for categorical variables. Postestimation tests, including Akaike information criteria (AIC) and the area under the receiver operating characteristic curve (i.e., C-statistic), were used to assess model fit. The *C*-statistic for our adjustment model for the BFS-PM was 0.5835 with an AIC of 36128.58. Scores are weighted as follows: First, all case-mix adjustor variables are entered into a logistic regression model, predicting a response of "excellent" on the BFS-PM at the patient level. A propensity score, or predicted probability, for an "excellent" response is derived from the results of the logistic regression. Finally, weights for the case-mix adjustment are calculated by taking the reciprocal of the propensity score and are applied all facility-level BFS outcomes. Weights are re-calculated quarterly.

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.

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?

Since social risk factors can potentially bias the results of performance measures (i.e. BFS scores), patient characteristics that are not under the provider's control, but are associated with the outcome of interest, should be accounted for using a case-mix adjustment. In the development of the Hospice CAHPS survey, CMS developed a preliminary case-mix adjustment model that includes factors such as decedent age and primary diagnosis and survey respondent's relationship to the decedent. Beyond the Hospice CAHPS survey measures, however, there is limited information on the adjustment of differences in case mix for other end-of-life surveys. Our goal was to align our case-mix adjustment on the BFS with adjustment factors utilized on already established surveys of end-of-life assessment. Additional variables were considered and were either not available (primary language spoken in the home, patient education level, and respondent age) or not used by other established survey measures (SHEP, CAHPS). In order to be considered for the case-mix model, variables must affect how he respondents rate the experience of their care, independent of providers (Elliot MN, Zaslavsky AM, Goldstein E, et al. Effects of survey mode, patient mix and nonresponse on CAHPS Hospital Survey scores. Health Services Research Journal. 2009;44(2 pt 1):501-518). All reported CAHPS survey risk-adjustment variables that were available in VA databases were used for the BFS case-mix adjustment model. VA data is pulled from the patient's electronic medical record which has <1% missing data. Decedent age is determined by government record as well as patient report. Patient diagnoses and comorbidities are determined by ICD9/ICD10 codes recorded in the medical record. Next of kin relationship is taken from the patient medical record and determined by both patient and next of kin. Mode of survey completion is recorded by research staff and/or survey administration vendor as soon as the BFS is completed by either mail, telephone or website. All adjustment variables are considered to have high integrity by industry standards and are used widely by VA investigators and leadership.

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)**

Kutney-Lee A, Carpenter J, Smith D, Thorpe J, Tudose A, Ersek M. (2018). Case-Mix Adjustment of the Bereaved Family Survey. American Journal of Hospice and Palliative Medicine, 1-8. doi: 10.1177/1049909117752669. In this analysis we found five critical risk adjustment factors: Veteran's age at the time of death (in years), number of medical comorbidities present at the time of death as defined by Van Walraven and colleagues' modification of the Elixhauser score, Veteran's primary diagnosis on last admission (classified into 1 of 15 clinical categories using the Agency for Healthcare Research and Quality Clinical Classification Software), relationship of Veteran's next-of-kin (i.e., spouse), and BFS administration mode (i.e., mail).

A social risk factor of patient comorbidity as defined by the Elixhauser score was determined to have an impact on the outcome developed because the facility patient comorbidity score is an indicator of the patient population that frequents the facility as well as the level of facility complexity (Acute care facility versus primarily nursing home facility, learning hospital, etc.) Social risk variables such as race and ethnicity were not included in the model because the BFS is used as a quality improvement measure. If you adjust for race/ethnicity (knowing that disparities exist) you are essentially letting facilities "off the hook" if they care for more racial/ethnic minorities.
Since social risk factors can potentially bias the results of performance measures (i.e. BFS scores), patient characteristics that are not under the provider's control, but are associated with the outcome of interest, should be accounted for using case-mix adjustment. Although our prior work has shown that race/ethnicity and region are associated with bereaved family assessments of quality of care, we choose not to include in our risk adjustment models for the BFS-PM. Our goal was to align our case-mix adjustment on the BFS with adjustment factors utilized in already established surveys of end-of-life assessment, specifically CAHPS-Hospice. Beyond the Hospice CAHPS survey measures, there is limited information on the adjustment of differences in case mix for other end-of-life surveys. Per CAHPS-Hospice risk-adjustment methodology for facility-level comparisons of performance scores (https://www.hcahpsonline.org/globalassets/hcahps/mode-patient-mix-adjustment/october_2019_pma_web_document.pdf), race/ethnicity and region are NOT included as adjustors. The CAHPS Hospice case-mix adjustment model includes the following variables: mode of survey administration, response percentile, decedent age, payer, primary diagnosis, length of hospice stay, caregiver age, caregiver education, caregiver relationship to decedent and language spoken at home. Further, our rationale for exclusion of these sociodemographic characteristics is also in alignment with AHRQ's recommendations for risk-adjusting performance scores for comparison purposes. AHRQ states that "most common risk adjustments—for age, prior medical history, or comorbidities—are considered a good idea. It is not considered appropriate, in most circumstances, to adjust for other sociodemographic characteristics such as race, ethnicity, income, education, and/or insurance status. Such adjustments would essentially bury information that could reveal what some would term unacceptable disparities in care." (https://www.ahrq.gov/talkingquality/translate/scores/adjustment-scoring.html).

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

To examine relationships between potential case-mix variables and the outcomes of interest, we constructed a set of regression models using logistic regression for the BFS-PM score. Models were fit using raw coefficients for categorical variables. In these models, age, Elixhauser comorbidity score, next-of-kin (child), primary diagnosis at death (neoplasm, mental illness, and diseases of the central nervous system/sense organs), and BFS response mode (telephone) were significantly associated with an "excellent" rating of the quality of care received by the veteran in the last month of life. Postestimation tests, including Akaike information criteria (AIC) and the area under the receiver operating characteristic curve (i.e., C-statistic), were used to assess model fit. The *C*-statistic for our adjustment model for the BFS-PM was 0.5835 with an AIC of 36128.58, and a Hosmer-Lemeshow goodness-of-fit test, p=0.1827.

Model variables, FY2013-FY2015	BFS-Performance Measure (range 0-1)[a]
- -	*OR (95% CI)*
Patient age in years [c]	1.02 (1.01-1.02)*
Patient Elixhauser Comorbidity score [c]	0.95 (0.93-0.96)*
Next-of-kin (NOK) relationship (BFS respondent)	- -
Spouse/partner	- -
Child	1.23 (1.15-1.32)*
Sibling	1.06 (0.98-1.15)
All other family members	1.05 (0.95-1.15)
Non-family member	0.99 (0.88-1.11)
Primary diagnosis at death	- -
Residual codes, unclassified, all E/V codes, none	- -
Infectious and parasitic diseases	0.92 (0.66-1.26)
Neoplasms	1.48 (1.09-2.01)*
Endocrine, nutritional and metabolic diseases & immunity disorders	1.24 (0.86-1.77)
Diseases of the blood and blood forming organs	1.21 (0.79-1.86)
Mental illness	1.67 (1.22-2.29)*
Diseases of the central nervous system and sense organs	1.49 (1.04-2.12)*
Diseases of the circulatory system	1.21 (0.89-1.66)
Diseases of the respiratory system	1.21 (0.88-1.66)
Diseases of the digestive system	1.07 (0.79-1.45)
Diseases of the genitourinary system	1.10 (0.80-1.51)
Diseases of the skin and subcutaneous tissue	0.75 (0.48-1.16)
Diseases of the musculoskeletal system and connective tissue	0.88 (0.60-1.29)
Injury and poisoning	0.91 (0.66-1.26)
Symptoms, signs and ill-defined conditions and factors influencing health status	1.20 (0.88-1.65)
Mode of survey administration	- -
Mail	- -
Telephone	0.78 (0.70-0.88)*
Online	0.93 (0.82-1.06)

[a] Logistic regression was used to test associations between BFS-Performance Measure scores and model variables. [b] Linear regression was used to test associations between Factor scores and model variables. [c] Coefficients have been standardized. *p≤0.05

- - cell intentionally left blank

```
--------------------------------------------------
Sensitivity                     Pr( +| D)   94.72%
Specificity                     Pr( -|~D)    8.83%
Positive predictive value       Pr( D| +)   62.80%
Negative predictive value       Pr(~D| -)   50.71%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   91.17%
False - rate for true D         Pr( -| D)    5.28%
False + rate for classified +   Pr(~D| +)   37.20%
False - rate for classified -   Pr( D| -)   49.29%
```

```
--------------------------------------------------
Correctly classified                  62.00%
--------------------------------------------------
```

Following adjustment, the BFS-PM scores of most facilities decreased by small amounts following case-mix adjustment. For example, nearly 40% of facilities experienced a positive or negative change of 0.5 percentage points or less on their adjusted BFS-PM score.

On average, facilities that benefited from case-mix adjustment cared for patients with greater comorbidity burden compared to facilities in the bottom 8% (5.9 vs 5.3 comorbid conditions). On average, the facilities whose scores increased most following adjustment had higher percentages of patients with a primary diagnosis of mental illness (14.5%) compared to the average hospital in the sample (9.7%) and to the facilities whose scores decreased by the greatest amount (7.2%).

The mean change in BFS-PM score across facilities was -0.6 with a range of -2.6 to 0.6. Overall, the BFS-PM scores of most facilities decreased following adjustment; however, the magnitude was small. The scores of 108 facilities decreased by less than 1 percentage point. Fifty-five facilities changed within +0.5 percentage points of their unadjusted score. The BFS-PM scores of 11 facilities increased as a result of case-mix adjustment. The national-level BFS Performance Measure score decreased by less than 1 percentage point.

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.**

On average, facilities that benefited from case-mix adjustment cared for patients with greater comorbidity burden compared to facilities in the bottom 8% (5.9 vs 5.3 comorbid conditions). On average, the facilities whose scores increased most following adjustment had higher percentages of patients with a primary diagnosis of mental illness (14.5%) compared to the average hospital in the sample (9.7%) and to the facilities whose scores decreased by the greatest amount (7.2%). Facilities that gained the most from case-mix adjustment were in urban areas in the Northwest and Midwestern regions of the country. Facilities that experienced the greatest decrease in BFS-PM scores tended to be located in rural areas in the Mountain and West regions of the United States.

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model <u>or</u> stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

To examine relationships between case-mix variables and the outcome of interest, we constructed a set of regression models using logistic regression for the BFS-PM score. Models were fit using both raw coefficients for categorical variables and standardized coefficients for continuous variables. Postestimation tests, including Akaike information criteria (AIC) and the area under the receiver operating characteristic curve (i.e., C-statistic), were used to assess model fit. Interaction terms between case-mix variables were tested but were not included in the final models because model fit was not improved. Facility-level scores were adjusted for case mix using inverse probability weighting. First, all case-mix adjustor variables were entered into a logistic regression model, predicting a response of "excellent" on the BFS-PM at the patient level. A propensity score, or predicted probability, for an "excellent" response was derived from the results of the logistic regression. Finally, weights for the case-mix adjustment were calculated by taking the reciprocal of the propensity score and were applied all facility-level BFS outcomes.

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 for our adjustment model for the BFS-PM was 0.5835 with an AIC of 36128.58.

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

Hosmer-Lemeshow goodness-of-fit test, p=0.1827

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

```
--------------------------------------------------
Sensitivity                    Pr( +| D)   94.72%
Specificity                    Pr( -|~D)    8.83%
Positive predictive value      Pr( D| +)   62.80%
Negative predictive value      Pr(~D| -)   50.71%
--------------------------------------------------
False + rate for true ~D       Pr( +|~D)   91.17%
False - rate for true D        Pr( -| D)    5.28%
False + rate for classified +  Pr(~D| +)   37.20%
False - rate for classified -  Pr( D| -)   49.29%
--------------------------------------------------
Correctly classified                       62.00%
--------------------------------------------------
```

2b3.9. Results of Risk Stratification Analysis:

The mean change in BFS-PM score across facilities was -0.6 with a range of -2.6 to 0.6. Overall, the BFS-PM scores of most facilities decreased following adjustment; however, the magnitude was small. The scores of 108 facilities decreased by less than 1 percentage point. Fifty-five facilities changed within +0.5 percentage points of their unadjusted score. The BFS-PM scores of 11 facilities increased as a result of case-mix adjustment. The national-level BFS Performance Measure score decreased by less than 1 percentage point.

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*)

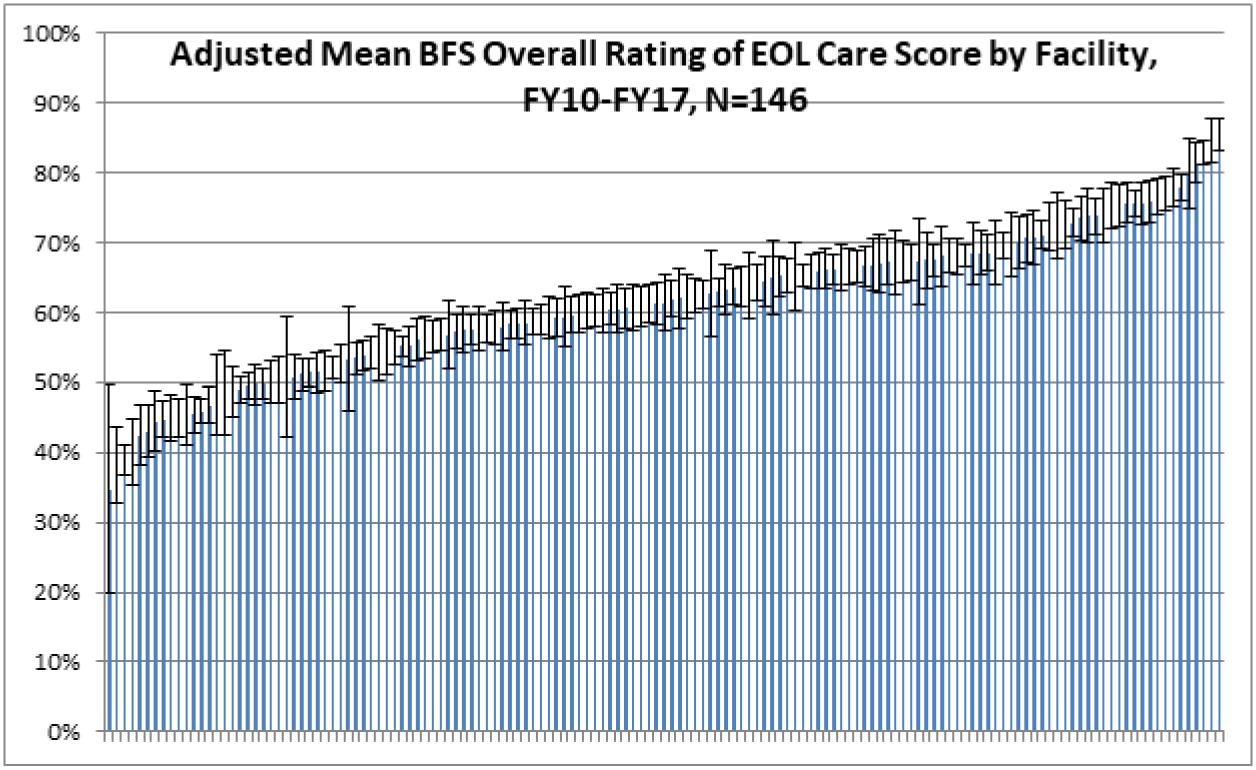
Using national VA medical records and survey data, we developed and tested a case-mix adjustment model for the BFS to facilitate valid comparisons of EOL care quality among 146 VA facilities. The BFS-PM scores of most facilities decreased by small amounts following case-mix adjustment. For example, nearly 40% of facilities experienced a positive or negative change of 0.5 percentage points or less on their adjusted BFS-PM score.

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

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

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

The facility-level score is the proportion of family members of deceased Veterans that rated overall end-of-life care as “Excellent”. All 146 facility scores are then averaged into one large national mean, weighted by the number of completed surveys in each facility.



Validity Analyses for the BFS Overall Rating of EOL Care Performance Measure:

Table 1 (below) provides descriptive statistics to demonstrate the variability in facility-level BFS Overall Rating of EOL Care PM and changes over time. Figure 2 depicts the range of mean Overall Rating of EOL Care facility scores for fiscal years 2010-2017, again demonstrating variability in scores (i.e., a performance gap).

Table 1. Adjusted Facility-level BFS Overall Rating of EOL Care[†] scores, FY10 – FY17 (n = 146 facilities)

Fiscal Year	Mean Overall Score[†]	Standard Deviation	Min/Max	Interquartile Range	Deciles
FY10[*]	57%	11	31/100	49 and 62	44, 47, 50, 52, 56, 58, 61, 65, 70
FY11[*]	58%	10	33/100	51 and 65	45, 49, 52, 55, 57, 61, 64, 66, 70
FY12[*]	59%	11	31/100	51 and 66	45, 50, 54, 57, 58, 62, 65, 67, 74
FY13[#]	63%	10	33/100	56 and 70	50, 55, 57, 61, 63, 66, 68, 71, 76

Fiscal Year	Mean Overall Score[†]	Standard Deviation	Min/Max	Interquartile Range	Deciles
FY14	62%	10	38/100	54 and 70	48, 51, 56, 59, 61, 64, 67, 71, 75
FY15	61%	11	22/100	53 and 71	48, 51, 55, 58, 61, 62, 66, 73, 77
FY16	63%	11	0/100	56 and 71	49, 53, 57, 60, 63, 68, 70, 71, 72
FY17	65%	11	13/100	58 and 72	52, 55, 58, 63, 66, 68, 69, 71, 72

[†] Family response to the question “Overall, how would you rate the care that [the Veteran] received in the last month of his life?” Dichotomized as Excellent v. all other responses [Very good, Good, Fair, Poor] (reported as %)

* During FY10-12 BFS was predominantly administered as a telephone survey

# During FY13 to present BFS was predominantly administered as a mailed survey

Several peer-reviewed publications have shown statistical and clinically meaningful differences in BFS performance:

Kutney-Lee A, Carpenter J, Smith D, Thorpe J, Tudose A, Ersek M. (2018). Case-Mix Adjustment of the Bereaved Family Survey. American Journal of Hospice and Palliative Medicine, 1-8. doi: 10.1177/1049909117752669. VAMC facility characteristics vary widely across the nation and those characteristics have an impact on BFS scores. To provide fair comparisons across facilities, the objective of this study was to develop a case-mix adjustment model for the BFS, and to examine changes in facility-level scores following adjustment. Following adjustment using model-based propensity weighting, the mean change in BFS-Performance Measure score across facilities was -0.6 with a range of -2.6 to 0.6. The scores of 108 facilities decreased by less than 1 percentage point, while 55 facilities changed within ±0.5 percentage points of their unadjusted score. On average, facilities that benefited most from adjustment cared for patients with greater comorbidity burden and were located in urban areas in the Northwest and Midwestern regions of the country.

Carpenter J, McDarby M, Smith D, Johnson M, Thorpe J, Ersek M. (2017). Associations between timing of palliative care consults and family evaluation of care for Veterans who die in a hospice/palliative care unit. Journal of Palliative Medicine, 20(7), 745-751. doi: 10.1089/jpm.2016.0477 After adjustment for patient and facility characteristics, family members of veterans whose first PCC occurred 91-180 days prior to death were more likely to rate overall care as “excellent” compared with those whose PCC occurred 0-7 days prior to death, 67.9 v. 62.1%, respectively (Adjusted Odds Ratio=1.37; 95% confidence interval (CI) 1.08-1.73). Earlier PCC is associated with greater family satisfaction with care. Strategies aimed at conducting PCC earlier in life limiting illness are needed.

Ersek M, Miller S, Wagner T, Thorpe, J, Smith, D, Levy C, Gidwani R, Faricy-Anderson K, Lorenz K, Kinosian B, Mor V. (2017). Association between aggressive care and bereaved families’ evaluation of end-of-life care for Veterans with non-small cell lung cancer who died in Veterans Affairs facilities. Cancer. doi: 10.1002/cncr.20700. Over 72% of Veterans had at least one episode of aggressive care and 31% received chemotherapy in the last 30 days of life. In all units except HPC, when patients experienced at least one episode of aggressive care, bereaved families rated care lower than when patients did not have any aggressive care. For patients dying in HPC units, the associations between overall ratings of care and two or more inpatient admissions or any episode of aggressive care were not statistically significant. Rates of aggressive care were not associated with age, and family ratings of care were similar for younger and older patients.

Kutney-Lee A, Smith D, Thorpe J, del Rosario C, Ibrahim S., Ersek, M. (2017). Race/Ethnicity and End-of-Life Care Among Veterans. Medical Care, 55(4) 342-351. doi: 10.1097/MLR.0000000000000637. Statistically significant differences were observed by race/ethnicity on 1 of the 4 end-of-life quality indicators: black Veterans were less likely than whites to receive a chaplain consult (77% vs. 79%; adjusted OR, 0.83, 95%CI, 0.73-0.94; p=0.004). Among the 15 Bereaved Family Survey items, less favorable outcomes were observed for black, Hispanic, and other racial/ethnic minorities on 12, 8 and 5 items, respectively. In comparison to whites, minority Veterans were less likely to report excellent overall care by the following odds ratios: 0.57 (95%CI, 0.53-0.61; p<0.001) for blacks, 0.85 (95%CI, 0.76-0.94; p=0.002) for Hispanics; and 0.83 (95%CI, 0.71-0.97; p=0.02) for other races/ethnicities. Our study found marked racial/ethnic disparities in the quality of end-of-life care in a national sample of Veterans who receive care in the equal-access VA healthcare system. Family perceptions are a critical component of evaluating equity and quality of care at the end of life.

Thorpe, J. M., Smith, D., Kuzla, N., Scott, L., & Ersek, M. (2016). Does Mode of Survey Administration Matter? Using Measurement Invariance to Validate the Mail and Telephone Versions of the Bereaved Family Survey. Journal of pain and symptom management, 51(3), 546-556. doi: 10.1016/j.jpainsymman.2015.11.006. To examine nonresponse bias for the mailed version of the Department of Veterans Affairs (VA) Bereaved Family Survey Performance Measure (BFS-PM) and evaluate the effect of nonresponse bias on facilities' BFS-PM scores we created a model to predict the likelihood of response based on patient and clinical characteristics. We then applied inverse probability weights to examine their effect on facilities' scores. We also evaluated facility performance before and after weighting for nonresponse vis-a-vis varying benchmarks. We received 8,912 surveys (45% response rate). The mean change in facility BFS-PM scores after weighting was -2%, (range: -10 to +11). The scores of 31% of facilities changed more than +/- 2%. The number of facilities meeting hypothetical benchmarks of 60, 70 and 80% also changed as a result of weighting for nonresponse.

Wachterman, M. W., Pilver, C., Smith, D., Ersek, M., Lipsitz, S.R., Keating, N.L. (2016). Quality of end-of-life care provided to patients with different serious illnesses." JAMA Internal Medicine,176(8),1095-1102. doi: 10.1001/jamainternmed.2016.1200. The adjusted proportion of patients receiving a palliative care consult was highest among cancer patients (69.8%) and dementia patients (61.5%), while less than half of patients with ESRD, cardiopulmonary failure, and frailty received such consults (P<.001). The adjusted proportion of patients dying in the ICU was lowest among cancer patients (16.1%) and dementia patients (11.0%), while about one-third of patients with frailty, ESRD, and cardiopulmonary failure died in the ICU (P<.001). The adjusted proportion of patients with a DNR order at death was highest among cancer patients (94.6%) and dementia patients (93.5%), compared to 87-88% for other conditions (P<.001). The adjusted proportion of family members reporting that care in the last month of life was "excellent" was highest among those of patients with cancer (57.7%) and dementia (59.2%) and lowest among patients with ESRD (53.8%) and cardiopulmonary failure (53.5%) (P<.001).

Ersek, M., Thorpe, J., Kim, H., Thomasson, A., Smith, D. (2015). Exploring End-of-Life Care in Veterans Affairs Community Living Centers. Journal of the American Geriatrics Society, 63(4), 644-650. doi: 10.1111/jgs.13348. Family evaluation of overall EOL care and quality of EOL care indicators for Veterans who died in CLCs were better than those of Veterans dying in acute and intensive care units, but were worse than those dying in hospice/palliative care units. Findings indicate that care in the CLC can be enhanced through the integration of palliative care practices. Future research should identify key elements of enhancing EOL care in nursing homes.

Sudore, R. L., Casarett, D., Smith, D., Richardson, D. M., & Ersek, M. (2014). Family involvement at the end-of-life and receipt of quality care. Journal of pain and symptom management, 48(6), 1108-

1116. doi: 10.1016/j.jpainsymman.2014.04.001. Most decedents (94.2%) had an involved surrogate. Veterans with involved surrogates were more likely than those without to have had a palliative consult, AOR 4.46 (95% CI; 4.03-4.93), a chaplain visit, AOR 1.20 (95% CI; 1.08-1.33), and a DNR order, AOR 4.81 (95% CI; 4.27-5.42). They were also more likely to die in a hospice/palliative care unit, OR 2.24 (95% CI; 1.96-2.56).

Ersek, M., Smith, D., Cannuscio, C., Richardson, D. M., & Moore, D. (2013). A nationwide study comparing end-of-life care for men and women veterans. Journal of palliative medicine, 16(7), 734-740. doi: 10.1089/jpm.2012.0537. Receipt of optimal end-of-life care did not differ significantly between women and men with respect to frequency of: discussion of treatment goals with a family member, receipt of palliative consult, bereavement contact, and chaplain contact with a family member. Family members of women were more likely than those of men to report that the overall care provided to the Veteran had been “excellent” (Adjusted proportions: 63% vs. 56%; OR=1.33; 95% CI 1.10-1.61; p=0.003).

Smith, D., Caragian, N., Kazlo, E., Bernstein, J., Richardson, D., & Casarett, D. (2011). Can we make reports of end-of-life care quality more consumer-focused? Results of a nationwide quality measurement program. Journal of palliative medicine, 14(3), 301-307. doi: 10.1089/jpm.2010.0321. Interviews were completed with family members for 3,897 of 7,110 patients (55%). Items showed an approximately 5-fold range of weights, indicating a wide variation in the importance that families placed on aspects of palliative care (low: pain management, weight¼0.54, 95% CI 0.38-0.70; /P/<0.001; high: providers were ‘‘kind, caring, and respectful: weight¼2.46, 95% CI 2.24-2.68; /P/<0.001). Weights were homogeneous across patient subgroups, and there were no significant changes in facilities’ quality rankings when weights were used. Both weighted and unweighted scores showed similar evidence of the impact of process measures.

Casarett, D., Johnson, M., Smith, D., & Richardson, D. (2011). The optimal delivery of palliative care: a national comparison of the outcomes of consultation teams vs inpatient units. Archives of Internal Medicine, 171(7), 649-655. doi:10.1001/archinternmed.2011.87. Interviews were completed with family members for 5901 of 9546 patients. Of these, 1873 received usual care, 1549 received a palliative care consultation, and 2479 received care in a palliative care unit. After nonresponse weighting and propensity score adjustment, families of patients who received a palliative care consultation were more likely than those who received usual care to report that the patient’s care in the last month of life had been “excellent” (adjusted proportions: 51% vs 46%; odds ratio [OR], 1.25; 95% confidence interval [CI], 1.02-1.55; *P*=.04). However, families of patients who received care in a palliative care unit were even more likely to report excellent care (adjusted proportions: 63% vs 53%; OR, 1.52; 95% CI, 1.25-1.85; *P*_.001).

Alici, Y., Smith, D., Lu, H. L., Bailey, A., Shreve, S., Rosenfeld, K., ... & Casarett, D. J. (2010). Families' perceptions of veterans' distress due to post-traumatic stress disorder-related symptoms at the end of life. Journal of pain and symptom management, 39(3), 507-514. doi: 10.1016/j.jpainsymman.2009.07.011. Seventeen percent of patients (89 of 524) were reported to have had PTSD-related symptoms in the last month of life. PTSD-related symptoms caused discomfort less often than pain did (mean frequency score 1.79 vs. 1.93; Wilcoxon sign rank test, P < 0.001) but more often than dyspnea did (mean severity score 1.79 vs. 1.73; Wilcoxon sign rank test, P < 0.001). Family members of patients with PTSD related symptoms reported less satisfaction overall with the care the patient received (mean score 48 vs. 62; rank sum test, P < 0.001). Patients who received a palliative care consult (n ¼ 49) had lower ratings of discomfort attributed to PTSD-related symptoms (mean 1.55 vs. 2.07; rank sum test, P ¼ 0.007).

Casarett, D., Smith, D., Breslin, S., & Richardson, D. (2010). Does Nonresponse Bias the Results of Retrospective Surveys of End-of-Life Care?. Journal of the American Geriatrics Society, 58(12), 2381-

2386. doi: 10.1111/j.1532-5415.2010.03175.x. After creating a model to predict the likelihood of response based on patient and clinical characteristics, we applied inverse probability weights to examine their effect on facilities' scores. We also evaluated facility performance before and after weighting for nonresponse vis-a-vis varying benchmarks. We received 8,912 surveys (45% response rate). The mean change in facility BFS-PM scores after weighting was -2%, (range: -10 to +11). The scores of 31% of facilities changed more than +/-2%. The number of facilities meeting hypothetical benchmarks of 60, 70 and 80% also changed as a result of weighting for nonresponse.

Lu, H., Trancik, E., Bailey, A., Ritchie, C., Rosenfeld, K., Shreve, S. ...& Casarett, D. (2010). Families' perceptions of end-of-life care in Veterans Affairs versus non-Veterans Affairs facilities. Journal of palliative medicine, 13(8), 991-996. doi: 10.1089/jpm.2010.0044. In bivariate analysis, patients who died in VA facilities (n=520) had higher mean satisfaction scores compared to those who died in non-VA facilities (n=89; 59 vs. 51; rank sum test *p*=0.002). After adjusting for medical center, the overall score was still significantly higher for those dying in VA facilities (*β*=0.07; CI=0.02-0.11; *p*=0.004).

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*)

See peer reviewed publications listed in section **2b4.1**

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?*)

The BFS has consistently shown the ability to identify statistically and clinically significant differences in performance across several entities including 1) receipt of a palliative consult results in higher BFS scores nationally and facility-wide; 2) death in an inpatient hospice unit (vs. other inpatient venues) consistently results in higher BFS scores nationally and facility-wide; 3) diagnoses at the end of life impact BFS results at the patient-level leading clinicians to implement quality improvement initiatives focusing on those diagnoses that consistently report lower scores ; 4) gender; 5) receipt of aggressive care results in lower BFS scores nationally and facility-wide; 6) chaplain interaction results in higher BFS scores nationally and facility-wide; 7) bereavement support after death results in higher BFS scores nationally and facility-wide; 8) DNR order at time of death results in higher BFS scores nationally and facility-wide; 9) race/ethnicity; 10) family involvement results in higher BFS scores nationally and facility-wide; and 11) earlier palliative consults results in higher BFS scores nationally and facility-wide.

____

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.***

The BFS methodology now uses one set of specifications across the VA system to adjust for risk and non-response. The survey questions are identical regardless of mode of administration (phone and mail). Therefore, mode of survey administration is included in the risk-adjustment model.
Please see 2a3 and 2b2 above.

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

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

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

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 non-responders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

To evaluate the effect of nonresponse on BFS-PM scores, associations were examined between several demographic and clinical characteristics and likelihood of survey response. All variables were considered for inclusion in a multivariable model in which the dependent variable was survey completion. This and all subsequent models used robust jackknife standard errors, clustered according to facility (146 clusters). After creating a model to predict the likelihood of response based on patient and clinical characteristics, we applied inverse probability weights to examine their effect on national and facility-level scores.

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*)

Rates of response by facility varied, with a mean response rate of 48% (*n*=146) and a range of 29% to 73%. Response rates for facilities with distinct characteristics were also wide-ranging. For example, facilities with an inpatient hospice unit had a higher rate of response when compared to facilities that did not have an inpatient hospice unit (mean response=49% (n=58); range=29%-69% vs. mean response=46% (n=88); range=29%-79%). Also, family members of patients who died in a facility that is categorized as “low complexity” (e.g., non-tertiary care facilities with a majority of non-acute care beds) were more likely to respond to the BFS than family members of patients who died in a “high complexity” facility (mean response=53% (n=60); range=29%-69% vs. mean response=44% (n=86); range=29%-79%). The mean change for all facilities before and after weighting for nonresponse was -2% points, with a range of -10 to +11. Of the

146 facilities in the sample, the scores of 45 facilities (31%) changed more than 2% points in either direction. Twenty-one facilities (14%) did not change their BFS-PM scores.

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 non-responders) 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*)

The mean change in facility BFS-PM scores after weighting was −2%, (range: −10% to+11%). The scores of 31% of facilities changed more than±2%. The number of facilities meeting hypothetical benchmarks of 60%, 70%, and 80% also changed as a result of weighting for nonresponse. The results underscore the importance of appropriately addressing nonresponse in the use of quality-of-care metrics based on Bereaved Family Survey (BFS) data.

3. Feasibility

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.

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), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

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**.

Some data elements are in defined fields in electronic sources

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).

The Bereaved Family Survey responses cannot be captured from electronic records because are the perceptions of family members of care of the Veteran.

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: Feasibility_Scorecard_v1.0.xlsx

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.

Three key lessons have emerged. First, procedures have been developed to permit an efficient and accurate identification of deaths as well as risk-adjustment and nonresponse bias adjustment variables from the EHR. These include a system of checks to ensure that veterans are deceased, and that they are eligible and that the data collected from the EHR is valid and accurate. Second, we have refined contact procedures to maximize interviewer/mailed survey efficiency, thereby decreasing costs. Third, we have developed operating procedures for addressing unresolved issues that are identified during interviews. This allows interviewers to make rapid referrals to the appropriate VA resources to provide assistance to bereaved family members (e.g. for assistant with burial or funeral benefits). We have also made several thousand referrals for bereavement/grief support to family members of deceased Veterans using these operating procedures.

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)*.

none

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 Quality Improvement (external benchmarking to organizations)	Quality Improvement (Internal to the specific organization) VA&acute;s Hospice and Palliative Care http://www.va.gov/geriatrics/Guide/LongTermCare/Hospice_and_Palliative_Care.asp

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

Program/Sponsor: Department of Veterans Affairs´ Hospice and Palliative Care Program Office.

Purpose: 1. to identify and reduce unwanted variation in the quality of end-of-life care throughout the VA health system. 2. to define and disseminate processes of care ("best practices") that contribute to improved outcomes for Veterans near the end-of-life and their families.

Geographic area: All 21 Veterans Integrated Service Networks (VISNs) and 146 inpatient facilities (acute and long-term care). Over 82,000 inpatient deaths have been reviewed for inclusion and for 87% these deaths, a survey was sent to the identified next of kin on record.

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?*)

VA is moving forward with a phased approach to empower terminally ill Veterans and their families in choosing between hospice care in a VA nursing home or in a community nursing home. Approximately 7,000 terminally ill Veterans are provided VA-paid hospice care in community nursing homes yearly however neither VA nor the Centers for Medicare and Medicaid provide quality reporting specific to the patient/family experience of end of life care in the community (non-VA) nursing home setting.

VA has begun administering the Bereaved Family Survey to the families of decedents that received VA-paid hospice care in community nursing homes to provide an important indicator of quality. These survey results will be disseminated to VA hospice care coordinators and facility leaders across VA to; 1) assist VA in selecting community nursing homes for Veterans' end of life care, 2) provide community comparisons on the quality of end of life care in VA nursing homes and 3) give Veterans and their families meaningful quality information to select the best venue for end of life care.

BFS is not currently used for consumer choice, but for transparency and quality improvement efforts.

BFS results are reported on an internal VA dashboard which promotes accountability

Developers plan to modify the scaling to allow for comparisons between BFS-PM and CAHPS-Hospice for community nursing home settings prior to the next submission.

BFS has been adopted by non-VA organizations. Stanford, Duke, UCLA, and Kaiser medical centers are all using the BFS to guide quality improvement efforts.

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.*)

n/a

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.

The Bereaved Family Survey results are reported to VA stakeholders, which include VA leadership, policy experts, clinicians, and researchers on a quarterly basis. The Veteran Experience Center along with the Hospice and Palliative Care Implementation Center are responsible for interpreting results of surveys directly with stakeholders on a regular/as-needed basis. Interventions based on survey results/data are them implemented specific to each stakeholder and then tracked over time to measure success. The end goal being an increase in the BFS global item score.

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.

Results are provided quarterly or on an as-needed basis for all stakeholders (defined above). BFS Performance Measure results at the national and facility-level are provided as well as additional process measure/quality indicators from the BFS itself and patient medical records. The VA Hospice and Palliative Care Implementation Center and VA Veteran Experience Center have regular contact with stakeholders for the intent of data interpretation and creation if quality improvement initiatives based on scores for the end goal of increasing satisfaction at the end of life. A Compendium of Best Practices SharePoint has been built to assist providers with quality initiatives and trend data is provided to track quality improvement initiatives impact on scores.

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.

Feedback is obtained on regular conference calls with stakeholders. Stakeholders have been very accepting of and open-to suggestions for increasing BFS performance measure scores.

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

Family members of deceased Veterans are the data source. However they are asked to assess the care that the deceased Veteran received at the end of life. The family members, for the most part, are very grateful for the survey in that they consistently complete the survey (48% response rate) and they communicate appreciation for asking for feedback. We also provide other services within the survey contact such as bereavement support and assistance with VA benefits.

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

VA leadership and policy makers are frequent users of our data. BFS performance measure scores are used to allocate funding to VA hospice and palliative programs throughout the nation. Scores are also used to measure and track high and low performers for institution of performance plan for meeting national benchmarks.

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.

We revised the BFS performance measure to include a nonresponse bias and risk adjustment so that facility comparison could be more accurate as well as future comparisons with non-VA entities.

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.

National patient-level improvement is demonstrated by increasing scores over time from 52% N=8,432 in 2008 to 64% N=10,899 in 2018. Facility scores vary over time with 79% N=116 out of 142 facilities having an increase in scores between 2008 and 2018. In 2018, 81 facilities (55%) scored above the national mean of 64%.

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.

The survey identifies areas for improvement in care within facilities in the VA. However, some facilities have a very small number of deaths and so caution should be taken when interpreting results since sample size may be too small to identify significant findings. Additionally, when contacting family members for survey completion, several issues were communicated to our Center staff, such as suicidal family members, need for bereavement services, need for information about VA benefits, etc. We were then able to assist family members with each of these issues.

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

Assisting family members of deceased Veterans with services, such as bereavement support, VA benefit information, and suicidal proclamations.

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)

2651 : CAHPS® Hospice Survey (experience with 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.

Survey items different as well as coding of items, Target group is also different, We are specifically looking at inpatient Veteran deaths, regardless of hospice use. Currently, the BFS is the only tool assessing end of life care in a VA inpatient setting. We believe that assessing all deaths, not just hospice deaths, is critical to the VA mission of improving care for all Veterans regardless of choice of level of care at death. We do see any negative impact to interpretability or burden of data collection.

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.)

NQF 2651 CAHPS Hospice Survey

Although the Bereaved Family Survey is in many ways similar to the CAHPS Hospice Survey, it provides information on a specific population (Veterans) and measures the quality of care provided a single health care system. Unlike the CAHPS-Hospice, the BFS provides a coherent measurement strategy that allows comparisons across systems of care and sites of death in a single health care system. This measure assesses the quality of care of the largest unified health care system in the United States and cares for more than 5 million patients annually. Because it is a unified health system, the VA is uniquely situated to make use of the quality data that can be easily and quickly disseminated. The BFS also measures satisfaction of care that are unique to a Veteran population (i.e., survivor and funeral benefits, PTSD). The population of Veterans and families that the VA serves is unique in several key respects:

1) Veterans and their families may face different challenges at the end of life than non-Veterans do. The costs of hospitalization are less likely to be relevant to non-VA populations.

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:** Bereaved_Family_Survey_data_collection_instrument.doc

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Department of Veterans Affairs / Hospice and Palliative Care

Co.2 Point of Contact: Scott, Shreve, scott.shreve@va.gov, 717-228-5946-

Co.3 Measure Developer if different from Measure Steward: Department of Veterans Affairs/ Hospice and Palliative Care / PROMISE Center

Co.4 Point of Contact: Hien, Lu, hien.lu@va.gov, 215-823-5800-7932

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.

An Expert panel review was used to refine item content and wording. The expert panel included: Therese Bernardo Cortez, RN MSN NP; Kimberly Kelley, LCSW; Carol Luhrs, MD; Paul Swerdlow, JD; Kathleen Bixby, BSN; Carla Anderson, MSN, RN; and Karyn Berlin, MSW

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2008

Ad.3 Month and Year of most recent revision: 10, 2013

Ad.4 What is your frequency for review/update of this measure? 3 years

Ad.5 When is the next scheduled review/update for this measure? 10, 2015

Ad.6 Copyright statement: This material is based upon work supported (or supported in part) by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, HSR&D. Use or publication of any materials used in the Bereaved Family Survey is prohibited.

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: